ROBUST MULTILINGUAL AUTOMATIC GENRE IDENTIFICATION IN TEXTS

Taja Kuzman Pungeršek

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Asst. Prof. Nikola Ljubešić, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Assoc. Prof. Tomaž Erjavec, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Assoc. Prof. Petra Kralj Novak, Member, Central European University, Vienna, Austria; Jožef Stefan Institute, Ljubljana, Slovenia
Dr. Tanja Samardžić, Member, University of Zürich, Zürich, Switzerland

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Taja Kuzman Pungeršek

ROBUST MULTILINGUAL AUTOMATIC GENRE IDENTIFICATION IN TEXTS

Doctoral Dissertation

ZANESLJIVA VEČJEZIČNA AVTOMATSKA IDENTIFIKACIJA ŽANRA V BESEDILIH

Doktorska disertacija

Supervisor: Asst. Prof. Nikola Ljubešić

Ljubljana, Slovenia, April 2026

Acknowledgments

First, I would like to thank my supervisor, Nikola Ljubešić, without whom this PhD thesis would not have been possible. Thank you for believing in my capabilities, for introducing me to the world of machine learning and natural language processing, and for your continuous support, motivation, and stream of research ideas. I also extend my sincere thanks to the members of the thesis evaluation board – Tomaž Erjavec, Petra Kralj Novak, and Tanja Samardžić – for their valuable feedback, which helped improve the thesis.

Second, the institutions that financially supported my research also deserve gratitude: the Department of Knowledge Technologies at the Jožef Stefan Institute, the CLARIN.SI infrastructure, and the Slovenian Research and Innovation Agency (ARIS) that funded the following research projects and programmes in which I was involved: the projects "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023), "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642), "Embeddings-based techniques for Media Monitoring Applications" (L2-50070, co-funded by Kliping d.o.o.) and the research programme "Language resources and technologies for Slovene" (P6-0411). My research was also supported by funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341 (the MaCoCu project), and from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751 (the ParlaCAP project). Please note that this communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

Third, I wish to thank everyone with whom I have collaborated and co-authored research papers, especially Peter Rupnik, Nikola Ljubešić, and Igor Mozetič – I have learned a great deal from you. I would like to thank my first mentors and collaborators who introduced me to the world of research during my undergraduate and Master's studies: Darja Fišer, Špela Vintar and Polona Gantar. I would also like to express my gratitude to fellow researchers working on the automatic genre identification task, especially Serge Sharoff, Veronika Laippala and her colleagues – our discussions on shared challenges and helpful exchanges about our work have significantly contributed to my research.

Fourth, I am grateful to my fellow PhD students, especially Katja Meden and Mojca Brglez, for sharing the challenges of the journey and for the strong mutual support. Mojca Brglez also deserves special thanks for all the hours spent on annotating the genre datasets – her dedication greatly contributed to the success of the GINCO annotation campaign. I would also like to thank my colleagues in the Department of Knowledge Technologies at the Jožef Stefan Institute for fostering a warm and welcoming academic environment – one that encourages the free exchange of ideas, mutual support, collaboration and innovation. Mili Bauer and Živa Antauer, thank you for handling all the administrative tasks that accompanied my research.

Finally, I would like to express my gratitude to my husband, family and friends for always believing in me and for providing me constant support and love. I am thankful to

my parents for encouraging me in all my pursuits and for being role models of hard work, self-motivation, and a relentless drive to grow and prosper – qualities that have shaped who I am today. I would also like to thank my friends and extended family for providing me with a rock-solid support system and for inspiring me with their example, proving that change and reinvention are always within reach when you believe in yourself. I am especially grateful to my husband for unconditional love, for always listening, motivating me to be my best self, and standing by me throughout my academic journey.

Dear reader, thank you for your interest in this work. I hope it proves useful to you.

Abstract

Collecting texts from the web has significantly accelerated the creation of large text datasets which are essential for the development of advanced language technologies, including large language models. However, because these texts are gathered automatically, their linguistic and functional characteristics are largely unknown, limiting their reliable use in research and applications. Automatic genre identification, a text classification task that categorizes texts into specific genre categories, provides key insights into such large-scale text collections and enables their filtering for applications in language technology and linguistic research.

This thesis advances automatic genre identification for web-scale multilingual text data through the creation of novel genre schemata, the development of manually-annotated genre datasets, and the exploration of robust machine learning methods. First, the study introduces new genre schemata designed to improve annotation reliability and cross-schema comparability, thereby addressing a key limitation of prior work where incompatible schemata hindered meaningful comparison across studies. We demonstrate that high-quality manual annotation with acceptable inter-annotator agreement can be achieved through a carefully designed genre schema, detailed guidelines, and the employment of expert annotators.

To address the lack of high-coverage genre datasets in our target languages, we develop high-quality, manually-annotated genre datasets in Slovenian and English, as well as multilingual test collections spanning eleven typologically diverse languages and different scripts. Building on the test datasets, we establish the AGILE benchmark, which enables standardized, reproducible cross-dataset and cross-lingual evaluation of genre classifiers. The benchmark also supports a systematic comparison of modern large language models across languages.

Using this infrastructure, we develop genre classifiers that achieve robust performance across various languages, datasets, and evaluation settings. To ascertain which machine learning methodology offers the most robust generalization, we experiment with a range of text classification techniques, ranging from traditional non-neural machine learning methodologies to cutting-edge approaches based on large language models. Our findings indicate that a BERT-like model, fine-tuned on our newly developed training dataset, achieves state-of-the-art performance across various datasets and languages, including languages using non-Latin scripts and languages not closely related to the fine-tuning languages. The resulting model is publicly released, providing a practical tool for large-scale automatic genre annotation of multilingual web text collections.

Furthermore, we propose the LLM Teacher-Student Framework, a novel approach for training text classifiers without manually-annotated data. By leveraging a large language model to generate training labels, this method enables scalable, cost-efficient development of genre classifiers, particularly for low-resource languages and specialized domains, and is applicable beyond genre identification to other text classification tasks.

Povzetek

Zbiranje besedil s spleta omogoča veliko hitrejše ustvarjanje velikih zbirk besedil, ki so nujno potrebne za razvoj naprednih jezikovnih tehnologij, med drugim tudi velikih jezikovnih modelov. A ker je ta metoda zbiranja besedil avtomatizirana, je malo znanega o jezikovnih in funkcijskih značilnostih besedil v zbirki, kar zmanjšuje njeno zanesljivost in uporabnost za nadaljnje raziskave in razvoj. Avtomatska identifikacija žanra – naloga kategorizacije besedil, ki besedila razvršča v določene kategorije žanrov – omogoča ključen vpogled v sestavo obsežnih zbirk besedil in njihovo filtriranje za razvoj jezikovnih virov in tehnologij ter za jezikoslovne raziskave.

V tej disertaciji prispevamo k napredku na področju razvoja modelov za avtomatsko identifikacijo žanra v obsežnih večjezičnih spletnih zbirkah besedil. Naše delo vključuje zasnovo novih shem kategorij žanrov, razvoj zbirk besedil, ročno označenih z žanri, in vrednotenje različnih metod strojnega učenja. Najprej predstavimo dve novi shemi kategorij žanrov, s katerima izboljšamo zanesljivost ročnega označevanja in primerljivost z drugimi obstoječimi shemami. Tako premostimo ključno pomanjkljivost predhodnih raziskav, ki temeljijo na neprimerljivih shemah in tako onemogočajo primerjavo rezultatov med deli. Pokažemo tudi, da je mogoče doseči kakovostno ročno označevanje s sprejemljivim strinjanjem med označevalci na podlagi premišljeno zasnovane sheme kategorij žanrov, podrobnih smernic za ročno označevanje in vključevanja visoko usposobljenih označevalcev.

Potem naslovimo pomanjkanje ustreznih učnih podatkov v naših ciljnih jezikih za to nalogo in razvijemo visokokakovostne ročno označene učne zbirke besedil v slovenščini in angleščini ter večjezične testne množice, ki zajemajo enajst jezikov iz raznovrstnih jezikovnih družin. Na podlagi testnih zbirk vzpostavimo ogrodje za vrednotenje AGILE, ki omogoča primerljivo in ponovljivo vrednotenje modelov za določanje žanra na različnih testnih podatkih in jezikih, poleg tega pa omogoča tudi sistematično primerjavo sposobnosti velikih jezikovnih modelov v različnih jezikih.

Sledi razvoj klasifikatorja, ki omogoča zanesljivo določanje žanra v različnih zbirkah besedil in jezikih. Da bi ugotovili, katera metoda strojnega učenja omogoča najboljše posploševanje, preizkusimo različne tehnike kategorizacije besedil, od klasičnih metod strojnega učenja do najsodobnejših pristopov, ki temeljijo na velikih jezikovnih modelih. Naše ugotovitve razkrijejo, da model tipa BERT, ki ga doučimo na naši novi učni zbirki besedil, doseže najboljše rezultate v različnih zbirkah besedil in jezikih in učinkovito določa žanre tudi pri besedilih, ki niso zapisana v latinici ali ki so zapisana v jezikih, ki niso sorodni jezikoma v učnih podatkih. Končni model javno objavimo in tako ponudimo prostodostopno orodje za avtomatsko označevanje žanrov v velikih večjezičnih zbirkah besedil.

Nenazadnje v disertaciji predlagamo novo metodologijo, ki omogoča razvoj klasifikatorjev besedil brez ročno označenih podatkov. Za označevanje učnih podatkov uporabimo velike jezikovne modele, kar omogoči hiter in učinkovit razvoj klasifikatorjev žanrov. Predlagana metodologija je lahko še posebej uporabna za tehnološko manj podprte jezike in za prilagajanje modelov na specifična področja, pristop pa se lahko prenese tudi na razvoj klasifikatorjev za druge naloge označevanja besedil.

Contents

List of Figures

List of Tables

Abbreviations

AGILE	...	Automatic Genre Identification Benchmark
API	...	application programming interface
BERT	...	bidirectional encoder representations from Transformers
CNN	...	convolutional neural network
CORE	...	Corpus of Online Registers of English
FAIR	...	Findable, Accessible, Interoperable and Reusable
FTD	...	Functional Text Dimensions
GINCO	...	Genre Identification Corpus
GPT	...	generative pretrained Transformer
LLM	...	large language model
LoRA	...	low-rank adaptation
NLP	...	natural language processing
SVM	...	support vector machine
TF-IDF	...	term frequency–inverse document frequency

Chapter 1

Introduction

In the introductory chapter, we first describe the problem addressed in the thesis (Section 1.1) and propose the goals we aim to achieve (Section 1.2). We present the initial hypotheses on which the thesis is based in Section 1.3. Finally, we conclude the chapter by presenting the scientific contributions the thesis offers in Section 1.4, outlining the methodology in Section 1.5, and providing an overview of the thesis structure in Section 1.6.

1.1 Background and Problem Definition

Collecting texts from the web has significantly accelerated the creation of large text collections, especially for less-resourced languages, such as Slovenian (Bañón, Esplà-Gomis, Forcada, García-Romero, et al., 2022). These text collections, often known as corpora, are essential for the development of advanced language technologies, including large language models, that rely on training with extensive datasets comprising millions of texts. To ensure a robust and accurate performance of large language models and other language technologies, it is crucial that the underlying massive collections of texts are of high quality (Penedo et al., 2023). However, due to automated collection methods, even corpora creators lack information on the characteristics of the texts that constitute these datasets (Baroni et al., 2009). Automatic genre identification, a text classification task that categorizes texts into specific genre categories, is the only feasible approach for determining the genre of individual texts within large-scale text collections. Specifically, this thesis focuses on automatic genre identification for large-scale web-based text collections. Enriching web collections with genre information would be particularly valuable, as they are widely used as training data for language technologies and as a focus of linguistic research due to their recency, massive size, and accessibility.

Genres can be defined as text categories, determined by the author's purpose, the typical function of the text, and its conventional form (Orlikowski & Yates, 1994). Multiple discourse, language and textual phenomena can be observed when studying the concept of genre. As a result, genre has been a subject of research across a range of disciplines, including linguistics, discourse studies, literary theory, media theory, as well as computational linguistics and information retrieval (Chandler, 1997). In this thesis, we explore automatic genre identification through methodologies derived from the field of language technology that focuses on the computational processing of human languages. This field integrates computational linguistics and natural language processing (NLP), thereby combining linguistics and computer science with a focus on artificial intelligence.

Automatic genre identification has been shown to be useful for various tasks. With the advent of the World Wide Web, it has become possible to query and collect unprece-

dented quantities of texts, leading to an increased interest in identification of genres in (web-based) text collections. Beyond the application of genre identification in the development and curation of web corpora, there has been large interest in this task within the field of information retrieval (Boese, 2005; Finn & Kushmerick, 2006; Priyatam et al., 2013; Roussinov et al., 2001; Stein et al., 2010; Stubbe & Ringlstetter, 2007; Vidulin et al., 2007; Zu Eissen & Stein, 2004). Integrating genre-based filtering into information retrieval tools would enable users to retrieve relevant texts more efficiently (Crowston et al., 2010). Additionally, leveraging genre information was shown to be beneficial for many other NLP tasks, including part-of-speech tagging (Giesbrecht & Evert, 2009), dependency parsing (Müller-Eberstein et al., 2021), automatic summarization (Stewart & Callan, 2009), machine translation (Van der Wees et al., 2018), and automated genre-aware assessment of the credibility of web documents (Agrawal et al., 2019).

As information on the genres of texts can be highly beneficial for various purposes, models for automatic genre identification would serve as a valuable tool, since manual annotation of millions of texts with genre labels is practically unfeasible. However, when approaching this task, researchers must confront several challenges related to defining the concept of genre, developing genre schemata, conducting manual annotation campaigns, and ensuring robust classifier performance. The key challenges encountered in the task of automatic genre identification can be summarized as follows:

1. Lack of consensus on the main concepts in automatic genre identification: There is no agreement on the fundamental concepts related to this task (Chandler, 1997), including on the term used to describe this phenomenon (D. Lee, 2002), leading to limited comparability among existing genre-annotated datasets (Sharoff et al., 2010).

2. Genre ambiguity in web-based texts: The development of the genre schema, as well as manual and automatic annotation need to address challenging cases arising from limited control over content creation on the web, such as texts comprising multiple intertwined genres or texts lacking any discernible characteristics of genre categories (Kwaśnik & Crowston, 2005; Repo et al., 2021; Sharoff, 2021). Additionally, as genre is considered as a text-level phenomenon (Biber & Conrad, 2019), identifying it in web corpora can be difficult, where texts might be only partially available due to imperfect automatic methods used for collection and cleaning of the datasets.

3. Challenging manual annotation: Ensuring high levels of agreement on genre labels among annotators during manual annotation campaigns has proven to be challenging (Egbert et al., 2015; Zu Eissen & Stein, 2004) or, according to Suchomel (2020), even "hardly possible". Discrepancies among annotators can result in unreliable genre datasets that are less appropriate for machine learning experiments (Sharoff, 2010). Additionally, in the absence of a reference genre schema, researchers typically developed their own, devised in accordance with the aims of their research (see for instance Asheghi et al. (2016), Berninger et al. (2008), Crowston et al. (2010), Egbert et al. (2015), Roussinov et al. (2001), Sharoff (2018), and Williams (2000)). This practice has resulted in a proliferation of schemata that are not comparable.

4. Issues with availability and coverage of existing genre datasets: Most datasets from previous studies are either hard to locate or no longer accessible, such as datasets from Dewe et al. (1998), Roussinov et al. (2001), M. Santini (2007), and Stubbe and Ringlstetter (2007). Additionally, genre research has predominantly focused on English, with a few studies conducted on Russian (Lepekhin & Sharoff, 2022; Sharoff, 2010, 2018, 2021), Finnish (Laippala et al., 2019; Skantsi & Laippala, 2023), Korean (Y.-B. Lee & Myaeng, 2002; Lim et al., 2005), and other languages (Laippala et al.,

2020; Maeda & Hayashi, 2009; Repo et al., 2021; Stamatatos et al., 2000). To the best of our knowledge, there exists no genre-annotated dataset for any South Slavic language.

5. Ensuring robustness of genre classifiers: Ensuring that genre classifiers perform well in out-of-domain scenarios rather than being overly reliant on specific datasets is a major challenge. In this task, traditional non-neural machine learning methods, such as support vector machines, have been shown to have limited capabilities of generalization beyond their training datasets (Sharoff et al., 2010). These issues might be mitigated using recently introduced deep neural models, such as pre-trained Transformer-based BERT-like language models that can achieve strong performance on this task on a small amount of training data, and were shown to be capable of good levels of cross-lingual transfer (Repo et al., 2021; Rönnqvist et al., 2021).

1.2 Purpose and Goals of the Dissertation

This thesis aims to address the complex task of automatic genre identification, specifically targeting automatic annotation of massive multilingual web corpora, by proposing novel methodologies for schema creation and manual annotation, and by exploring appropriate machine learning techniques. As we aim to provide automatic genre identification solutions for massive multilingual web text collections, our research is based on the three important criteria for the developed genre classifier: 1) robustness, 2) multilinguality, and 3) high usability for downstream tasks.

In our work, we define robustness as the high capabilities of the model to generalize across datasets. The developed model will be useful for automatic annotation of web texts only if it achieves high prediction performance on the newly provided web text collections, not only on the datasets on which it was trained. Secondly, to be able to reliably evaluate the model's performance, it is crucial that we also achieve robustness of the manually-annotated genre datasets. We consider the datasets to be robust if the manual annotations are reliable. Reliability of the manually-annotated datasets is evaluated based on inter-annotator agreement, where the annotated labels are considered to be trustworthy if the annotators mostly agree on the labels.

Multilinguality is one of the fundamental values in our research, as we seek to avoid the exclusive focus on English that has been studied in most prior research. In addition to catering to English, we aim to provide genre datasets and classifiers also for Slovenian and other less-resourced European languages. That is why robustness will also be evaluated based on the capabilities of the model to generalize across various languages.

Additionally, as we aim to develop models for automatic genre classification that would enable automatic annotation of massive web collections, we set forward an additional criterion related to the usability of the proposed approach. Usability should be especially considered during the development of the genre schema. Firstly, the genre schema that the model uses should provide sufficient coverage that the majority of texts in the web text collection would be assigned a specific label. At the same time, it is advisable that the genre set comprises a limited number of labels, as larger sets result in an increased computation cost due to a larger final layer of the neural model and a bigger size of the output tensors. Secondly, we aim to ensure that the genre-enriched web corpora would be useful for language technology developers as well as corpus linguists who would extract or filter texts based on genre information. This means that the names of the proposed genre labels should be recognizable to common users as much as possible.

Furthermore, we are committed to adhere to the principles of findability, accessibility, interoperability, and reusability (FAIR) (Wilkinson et al., 2016) throughout our research.

The datasets and developed classifiers will be made publicly accessible through data repositories to ensure their long-term availability. This initiative aims to promote further research in this area and to enhance the long-term usefulness of the contributions of our work.

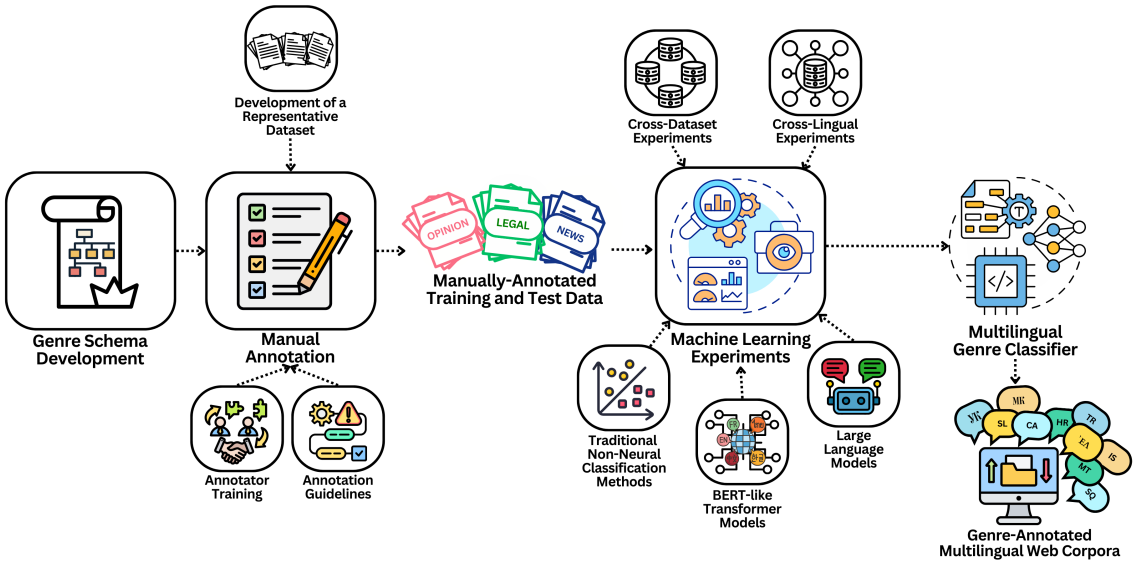


Figure 1.1: An overview of the phases of automatic genre identification research addressed in this thesis, including the construction of a genre schema, the creation of manually-annotated training and test datasets, machine learning experiments leading to the development of a multilingual genre classifier, and the application of this classifier to multilingual web corpora.

We aim to address all stages of the development of automatic genre classifiers, as shown in Figure 1.1. These stages include constructing a genre schema, creating a manually-annotated dataset, conducting machine learning experiments, validating classifier robustness, and applying the developed classifiers to massive web corpora. Accordingly, this thesis aims to achieve the following goals:

G1. Develop a high-coverage genre schema that encompasses the full spectrum of genre variation in web texts and ensures reliable manual annotation with high inter-annotator agreement.

G2. Propose a novel methodology for manual genre annotation that ensures high reliability based on a novel genre schema, detailed annotation guidelines, and continuous training of expert annotators.

G3. Develop a genre dataset for Slovenian, which will be the first of its kind for this language, to enable training and evaluation of machine learning models in automatic genre identification on Slovenian language. Additionally, we will construct training datasets that include English, and separate test datasets for various European languages to assess the performance of genre classifiers in cross-dataset and cross-lingual evaluation scenarios.

G4. Establish a benchmark for cross-dataset and cross-lingual assessment of machine learning approaches in automatic genre identification, based on newly developed test datasets in English, Slovenian, and various other European languages. With the benchmark, we aim to address the current lack of a reference dataset and enable

comparative analyses of machine learning techniques employed in the task of automatic genre identification.

G5. Develop a robust multilingual genre classifier capable of generalizing across datasets and European languages by exploring various machine learning approaches, including deep neural language models. Additionally, we will experiment with training on multiple genre datasets to enhance performance and robustness, evaluated in cross-dataset and cross-lingual experiments.

1.3 Hypotheses

Previous work showed that manual annotation for automatic genre identification is a very challenging task (Egbert et al., 2015; Sharoff, 2018; Suchomel, 2020; Zu Eissen & Stein, 2004). Research that evaluated annotation reliability based on the inter-annotator agreement, using Krippendorff's alpha (Krippendorff, 2018) as a metric, mostly did not reach the threshold value of 0.667, which is established as the minimum acceptable level by Krippendorff (2018). The low level of annotation agreement has negatively impacted the reliability of existing genre datasets (Sharoff, 2010). This thesis aims to improve robustness of the manually-annotated datasets by proposing novel genre schemata and annotation methodologies to ensure reliable manual annotation. We define the manual annotation as reliable if the inter-annotator agreement is acceptable, meaning that it is above the threshold Krippendorff's alpha value of 0.667. This leads to our first hypothesis:

H1. Acceptable inter-annotator agreement in genre identification annotation campaigns can be accomplished through a manual annotation approach that is based on a well-designed genre schema, comprehensive annotation guidelines, and employment of expert annotators.

Recent previous research has demonstrated promising capabilities of pre-trained Transformer-based language models for the task of automatic genre identification (Repo et al., 2021; Rönnqvist et al., 2021). Building on these findings, we will conduct machine learning experiments with various monolingual and multilingual pre-trained Transformer-based models to evaluate their performance on this task specifically when applied to Slovenian. These models will also be compared with other machine learning methods commonly used for genre identification and other text classification tasks to identify the machine learning method that is the most appropriate for our task. Following this, we propose the second hypothesis:

H2. Fine-tuned deep neural Transformer-based language models outperform other machine learning methods in the task of automatic genre identification in the Slovenian language.

Improving cross-lingual performance would greatly increase the usefulness of genre classifiers, as it would enable their deployment across various languages without the need for manual annotation. Recent research has demonstrated promising results related to the cross-lingual capabilities of Transformer-based genre classifiers (Repo et al., 2021; Rönnqvist et al., 2021). To address the needs of numerous less-resourced European languages, we aim to research the optimal approaches to development of genre classifiers capable of achieving high zero-shot cross-lingual performance in a broad set of languages. We aim to develop multilingual models that achieve robust performance on various languages, that is, that are capable of generalizing not only across datasets but also languages. Based on

model performance achieved in prior work (Repo et al. (2021), Rönnqvist et al. (2021), and Sharoff (2021); see Table 2.3 in Section 2.3) and human performance on this task, measured by the inter-annotator agreement, we consider micro-F1 and macro-F1 scores of around 0.70 as the minimum threshold for acceptable performance of a genre classifier. We deem the classifiers to be reliable if they achieve micro-F1 and macro-F1 scores on the test dataset above 0.80. In relation to this goal, we propose the following hypothesis:

H3. A reliable zero-shot cross-lingual performance in automatic genre identification across numerous European languages is possible by leveraging the cross-lingual capabilities of massively multilingual pre-trained Transformer models.

1.4 Scientific Contributions

The thesis aims to offer valuable contributions to the research community, particularly in the domains of genre classification, text classification, multilingual natural language processing (NLP), and the development of language resources for less-resourced languages. The scientific contributions of this thesis are as follows:

1. Novel comprehensive genre schemata, developed with focus on capturing the diversity of genres found in web texts. We propose the GINCO schema that aims to improve upon existing schemata by addressing issues related to low inter-annotator agreement, and thus improving annotation reliability (see Section 3.1.1). We then further refine the GINCO label granularity and improve its level of comparability to other common high-coverage schemata, which leads to the development of the X-GENRE schema in Section 3.2.1. Crucially, the proposed schema is designed to be mappable to widely used earlier schemata, enabling the reuse of existing annotated datasets and systematic cross-lingual and cross-dataset comparisons. This directly addresses a major limitation of prior work, where incompatible genre schemata prevented a meaningful comparison of results across studies. The novel schemata are described in the papers "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al. (2022); Section 3.1.5), and "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al. (2023); Section 3.2.4).

2. A novel methodology for manual genre annotation designed to ensure high inter-annotator agreement and improve the reliability of manually-annotated datasets. The precision and consistency of the manual annotations is improved through the application of objective annotation criteria and the employment of expert annotators, supported by continuous training and detailed guidelines that are made publicly available to support future manual annotation campaigns.[1] The details on the annotation methodology are provided in Section 3.1.2 that summarizes the findings from the paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al. (2022); Section 3.1.5).

3. Multiple manually-annotated genre datasets for Slovenian, English, and various other less-resourced European languages. The thesis addresses a gap in genre classification research by introducing several high-quality, manually-annotated genre datasets

[1]The GINCO annotation guidelines are available at https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/, and the X-GENRE annotation guidelines are published at https://tajakuzman.github.io/X-GENRE-Annotation-Guidelines/.

for languages that have been previously excluded from studies on automatic genre identification, namely the GINCO dataset for Slovenian (Kuzman et al. (2021); see Section 3.1.3 and the paper by Kuzman, Rupnik, et al. (2022) in Section 3.1.5), the English-Slovenian X-GENRE dataset (Kuzman and Ljubešić (2024a); see Section 3.2.2 and the paper by Kuzman, Mozetič, et al. (2023) in Section 3.2.4), the English EN-GINCO test dataset and the X-GINCO test dataset in ten European languages (see Section 4.1.1, and the papers by Kuzman, Mozetič, et al. (2023) and Kuzman Pungeršek et al. (In submission) in Sections 3.2.4 and 4.6). These datasets enable the development and evaluation of multilingual genre classifiers, as shown in Section 3.2 and Chapter 4. Additionally, the multilingual test datasets, integrated into the AGILE Benchmark (Automatic Genre Identification Benchmark), freely available on GitHub,[2] enable a comprehensive evaluation of genre classifiers – as well as an evaluation of broader capabilities of large language models – across languages and scripts.

4. A novel genre classifier that demonstrates robust performance across languages and datasets, with a particular focus on less-resourced languages, including Slovenian. The research systematically compares multiple machine learning approaches for text classification to identify the most effective methods for automatic genre identification in diverse European languages, described in Sections 3.1.4, 4.2, 4.3, 4.4, and 4.5, and in papers "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al. (2022); Section 3.1.5), "*Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments*" (Kuzman, Ljubešić, and Pollak (2022); Section 3.2.3), "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al. (2023); Section 3.2.4), "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al. (In submission); Section 4.6), and "*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?*" (Kuzman Pungeršek, Rupnik, Porupski, et al. (2026); Section 4.7). We evaluate models in monolingual, multilingual, and zero-shot cross-lingual settings, revealing that high performance can be achieved even on diverse and unrelated languages using different scripts. These experiments provide new evidence of the robustness and transferability of genre classifiers across languages and datasets.

5. A novel methodology for developing text classifiers without the need for manually-annotated data, termed the LLM Teacher-Student Framework. We show that substituting human annotators with a large language model for training data annotation is an effective strategy for developing robust and efficient genre classifiers, as well as models for topic classification. To the best of our knowledge, this methodology has not previously been applied to this task and is potentially transferable to a wide range of text classification tasks. The methodology is presented in Chapter 5 and published in the paper *"LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification"* (Kuzman and Ljubešić (2025); Section 5.3).

1.5 Methodology

In this thesis, we aim to address the complex task of automatic genre identification through the creation of novel genre schemata, the development of manually-annotated

[2]https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

genre datasets, and the exploration of robust machine learning methods. To fulfill the objectives and assess the hypotheses presented in the previous sections, the following methodology is employed:

1. The methodology for developing a high-coverage genre schema: We perform a thorough analysis of previously proposed genre schemata to analyze their similarities and differences. Based on this analysis, we develop a high-coverage genre schema that incorporates elements from 15 previous schemata (Boese, 2005; Dewe et al., 1998; Egbert et al., 2015; Laippala et al., 2019; Y.-B. Lee & Myaeng, 2002; Lim et al., 2005; Roussinov et al., 2001; M. Santini, 2007, 2010; Sharoff, 2010, 2018; Stamatatos et al., 2000; Stubbe & Ringlstetter, 2007; Vidulin et al., 2007; Zu Eissen & Stein, 2004). Previous machine-learning experiments have been limited by their dependency on specific datasets and schemata (Rehm et al., 2008). In response, this thesis aims to explore the potential of mapping our proposed schema onto existing high-coverage schemata to enable cross-dataset experiments. Specifically, we examine the comparability of our schema with two commonly used high-coverage schemata: the schema of the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the schema of the Functional Text Dimensions (FTD) approach (Sharoff, 2018).

2. The methodology for constructing genre datasets: As the thesis focuses on the automatic identification of genres within web text collections, novel genre datasets are derived from existing web corpora, such as the Slovenian slWaC 2.0 corpus (Erjavec & Ljubešić, 2014), the MaCoCu web corpora (Bañón, Esplà-Gomis, Forcada, García-Romero, et al., 2022) in various European languages, and the English web corpus enTenTen20 (Jakubíček et al., 2013). The proposed methodology involves the extraction of random texts from web corpora, which also include noise, texts with multiple intertwined genres, and other web-specific challenges, enabling evaluation of classifiers under realistic conditions. The finalized datasets are made publicly available in data repositories and are formatted in widely recognized data formats to ensure maximum accessibility and reproducibility.

3. The methodology for manual annotation: A manual annotation campaign is conducted based on a methodology that ensures high inter-annotator agreement for the task of genre annotation. Unlike annotation campaigns that use crowdsourcing, we employ experts from the fields of linguistics and computational linguistics, who receive fair compensation. To ensure reliability, annotators are guided through the manual annotation process using a survey structured as a decision tree, with criteria for genre categories being as objective as possible. Continuous training is provided to annotators to ensure a good understanding of the task. The reliability of manual annotation is evaluated by calculating the inter-annotator agreement using Krippendorff's alpha (Krippendorff, 2018), a widely used metric that evaluates the annotator agreement while accounting for chance agreement. We aim to achieve a Krippendorff's alpha value above 0.667, which is considered the minimum acceptable threshold for reliable annotation (Krippendorff, 2018).

4. The methodology for machine learning experiments: The thesis aims to evaluate a range of machine learning methodologies commonly used in text classification, specifically focusing on the task of automatic genre identification. The machine learning experiments include both non-neural and neural models. The non-neural models include the Naive Bayes classifier, logistic regression classifier, and support vector machine (SVM), using the TF-IDF (term frequency–inverse document frequency) text representations. The neural models comprise the shallow neural fastText model

(Joulin et al., 2017), as well as a selection of monolingual and multilingual deep neural BERT (Bidirectional Encoder Representations from Transformers) models. Furthermore, we explore a promising novel approach involving the classification of texts using instruction-tuned decoder-only large language models (LLMs) in a zero-shot setting, where no training data is needed. Once we identify the most effective machine learning approach for both monolingual and multilingual genre classification, we extend the evaluation to cross-dataset and cross-lingual experiments to evaluate the model's robustness across various datasets and languages. The models' performance is assessed using metrics that are commonly used for single-label multi-class text classification, namely micro-F1 and macro-F1. These metrics provide insights into the models' performance at both the instance and label levels, respectively. Based on model performance achieved in prior work (Repo et al. (2021), Rönnqvist et al. (2021), and Sharoff (2021); see Table 2.3 in Section 2.3) and human performance on this task, measured by inter-annotator agreement, we consider micro-F1 and macro-F1 scores of around 0.70 as the minimum threshold for acceptable performance of a genre classifier, and performance above 0.80 as reliable model performance.

1.6 Organization of the Thesis

The thesis is structured as follows. Chapter 2 provides an overview of the main approaches, challenges and solutions identified in previous work, related to the definition of the genre concept and the genre schema (Section 2.1), the methods for dataset collection and annotation (Section 2.2), and the development of robust machine learning models (Section 2.3). The content of this chapter is based on the accompanying journal paper "*Automatic Genre Identification: A Survey*" (Kuzman & Ljubešić, 2023), enclosed in Section 2.4. The chapter concludes with Section 2.5 which identifies the gaps in prior work that are addressed in the remainder of the thesis.

Chapter 3 outlines the development of novel genre schemata, manually-annotated datasets, and genre classifiers as part of our goal to develop a robust multilingual genre classifier. In Sections 3.1 and 3.2.1, we introduce the newly developed GINCO and X-GENRE genre schemata, which address Goal *G1*. In Section 3.1.2, we present the annotation campaign and annotation guidelines for the GINCO schema, fulfilling Goal *G2* and providing evidence for Hypothesis *H1*. Finally, in Sections 3.1.3 and 3.2.2, we present the Slovenian manually-annotated GINCO dataset and the Slovenian-English manually-annotated X-GENRE dataset, respectively, fulfilling the first part of Goal *G3*. This chapter is based on the following accompanying papers: "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al., 2022) in Section 3.1.5; "*Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments*" by Kuzman, Ljubešić, and Pollak (2022) in Section 3.2.3; and "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4.

In Chapter 4, we conduct experiments with various text classification methods to determine which approach offers the most robust generalization across different datasets and languages. In Section 4.1, we detail the methodology and introduce the newly developed test datasets designed for benchmarking classification methods on automatic genre identification, addressing the second part of Goal *G3*. Building on these datasets, we establish a publicly available benchmark that fulfills Goal *G4*. The results of experiments assessing both in-dataset and cross-dataset performance on Slovenian and English are presented in Section 4.2 – where the results support Hypothesis *H2* – and in Section 4.3. We then

extend the evaluation to examine zero-shot cross-lingual performance in Section 4.4, where we provide empirical support for Hypothesis *H3*. Additionally, we assess the zero-shot performance of large language models (LLMs) on the task of automatic genre identification in Section 4.5. These machine learning evaluations enable us to identify the methodology that yields the best-performing genre classifier, addressing Goal *G5*. This chapter is based on the following papers: "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4; "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al., In submission) in Section 4.6; and "*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?*" (Kuzman Pungeršek, Rupnik, Porupski, et al., 2026) in Section 4.7.

In Chapter 5, we introduce a novel methodology for developing genre classifiers without the need for manually-annotated training data. Section 5.1 evaluates the reliability of an LLM when employed as a data annotator within the proposed LLM Teacher Student Framework approach. Subsequently, in Section 5.2, we assess whether the proposed methodology can produce reliable genre classifiers. This chapter is based on the paper *"LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification"* by Kuzman and Ljubešić (2025), enclosed in Section 5.3.

We conclude the thesis in Chapter 6 where we first summarize the research and then highlight its contributions in Section 6.1. Secondly, we outline the planned future work in Section 6.2. Lastly, to facilitate the replicability of our research, we provide details regarding the availability of the code used in our experiments, as well as access to the developed datasets and models in Section 6.3.

Chapter 2

Challenges and Proposed Solutions in Related Work

Although numerous studies over the past twenty years have aimed to develop efficient genre classifiers, at the start of this PhD research, no freely available genre classifier existed that could be applied to large-scale web text collections. Entering this field, one quickly discovers a lack of consensus on best practices across all stages of developing automatic genre identification models, requiring careful consideration and decision-making at each step. In this chapter, we present the main approaches, challenges and solutions identified in previous work, based on the detailed overview provided in the accompanying journal paper "*Automatic Genre Identification: A Survey*" (Kuzman & Ljubešić, 2023), enclosed in Section 2.4. Following the main goals of the thesis, the chapter is especially focused on robustness, multilinguality, high coverage, and usability of genre datasets and classifiers for downstream applications. It outlines the challenges related to the definition of the concept of genre and the genre schema (Section 2.1), the dataset collection and annotation methods (Section 2.2), and the development of robust genre classifiers (Section 2.3).

2.1 Definition of Genre Schema

When tackling the task of automatic genre identification, we must first decide which term to use to describe this phenomenon. There is little agreement among different groups of researchers on the fundamental concepts related to the task of automatic genre identification (Chandler, 1997), starting with the terminology used to describe the task (D. Lee, 2002). Previous work uses various terms to describe the concept of genre. In this work, we adopt the commonly used terms "genre", "automatic genre identification" and "web genre identification" (used by Asheghi et al. (2016), Crowston et al. (2010), M. Santini (2007), Stubbe and Ringlstetter (2007), Vidulin et al. (2007), and Zu Eissen and Stein (2004), among others). However, some research groups working on the same task use alternative terms, such as "registers" (see Egbert et al. (2015) and further research by Laippala et al. (2019), Repo et al. (2021), and Rönnqvist et al. (2021)), "functional text dimensions" (Sharoff, 2018) or "text types" (Stubbs, 1996).

Secondly, previous work uses various criteria to define genre categories. Most definitions describe genre categories based on the socially recognized form and the intended communicative purpose (function) of the text (Kwaśnik & Crowston, 2005). Some definitions also take into account the target audience (Zu Eissen & Stein, 2004), reader expectations (S. M. Santini, 2006), style (Argamon et al., 1998; Finn & Kushmerick, 2006), or content (Rosso, 2008). To identify genres, researchers either follow the text-internal perspective, examining linguistic and other visible elements in the text, the text-external or functional

perspective, focusing on the function of the texts (see Sharoff (2010, 2021) for a detailed discussion on both perspectives), or a combination of both, as is done in our work.

In prior work, researchers defined genre categories differently and mostly developed their own genre schema in accordance with the aims of their research. This practice led to very limited comparability among existing genre schemata and datasets (Sharoff et al., 2010). Further insight into this issue comes from an analysis of 16 previously used genre schemata, specifically those presented in the works of Asheghi et al. (2016), Boese (2005), Dewe et al. (1998), Egbert et al. (2015), Kuzman, Rupnik, et al. (2022), Laippala et al. (2019), Y.-B. Lee and Myaeng (2002), Lim et al. (2005), M. Santini (2007, 2010), Sharoff (2010, 2018), Stubbe and Ringlstetter (2007), Suchomel (2020), Vidulin et al. (2007), and Zu Eissen and Stein (2004). Table 2.1 shows the frequency of genre labels that are shared by at least three schemata, highlighting the large variation among genre schemata. In total, they use 280 genre categories. Almost half of the labels (129) are used by only one schema. Only 20 labels are shared across multiple schemata, as shown in Table 2.1, with the most frequent category present in 9 out of 16 schemata. A further analysis of examples from existing genre datasets, included in the journal paper that is enclosed in Section 2.4, reveals an even more complex reality. Even when schemata share identical or similar label names, the labels do not necessarily correspond to similar texts, as their interpretation varies across schema authors and annotators. This variation further obscures the level of comparability between existing genre datasets. Therefore, attempts to merge genre schemata or datasets for model training or cross-dataset evaluation require a detailed examination of the datasets and their instances, rather than merging the datasets and schemata solely based on label names.

Table 2.1: Most frequent genre labels from 16 genre schemata. Labels with very similar names, e.g., “FAQ” and “FAQs” or “news” and “news/reporting”, were counted as the same label name.

Genre label	Occurrences in schemata
FAQ	9
Instruction	7
News	6
Legal	6
Personal Home Page	6
Review	6
Information	5
Discussion	5
Research Article	5
Interview	5
Lyrical	5
Poem	4
Recipe	4
E-shop	4
Editorial	4
Promotion	3
Link Collection	3
Journalistic	3
Blog	3
Prose	3

One of the main reasons for the significant differences between genre schemata is that they were constructed based on the aim of the research. Accordingly, the previously developed schemata can be divided into the following groups: high-coverage schemata, information retrieval schemata, and schemata, created with other aims. Information retrieval schemata comprise a selection of genre categories that are deemed useful for search engine users (see Dewe et al. (1998), Lim et al. (2005), Roussinov et al. (2001), M. Santini (2007), Vidulin et al. (2007), and Zu Eissen and Stein (2004)). Other schemata and corresponding genre datasets were developed to study specific genres, such as poetry and prose (Shavrina, 2019). Since our work aims to develop schemata and genre classifiers capable of efficiently capturing the full variation of web content, we focus on high-coverage schemata. They aim to cover the full range of genre variation on the web in order to annotate all web texts with a certain label. In accordance with this goal, they mostly use larger sets of specific classes (see Egbert et al. (2015), Laippala et al. (2019), Sharoff (2018), Stubbe and Ringlstetter (2007), and Suchomel (2020)), a smaller set of broader categories (see Sharoff (2010)) or a combination of the two (see M. Santini (2010)). Higher label granularity typically complicates the manual annotation process and can result in lower inter-annotator agreement and a less reliable manually-annotated dataset (Sharoff, 2010). On the other hand, using a smaller set of labels with categories that are too broad and consequently not separated well can negatively impact automatic genre identification (Asheghi et al., 2016).

The high-coverage schemata introduced in previous work provide a useful starting point, especially those by Berninger et al. (2008), Egbert et al. (2015), M. Santini (2010), Sharoff (2010, 2018), and Stubbe and Ringlstetter (2007). However, each of them has limitations that make it unsuitable for direct use in our study. Some have a too high label granularity, with over 40 genre labels, which has been shown to negatively impact manual annotation reliability (e.g., schemata by Berninger et al. (2008) and Egbert et al. (2015)). Others have a limited comparability to other existing schemata due to differences in label sets (e.g., schemata by M. Santini (2010) and Sharoff (2010, 2018)), or include abstract label names that would not be understood by end users, such as *resources* (Stubbe & Ringlstetter, 2007), *recreation* (Sharoff, 2010), and *commpuff* (Sharoff, 2018). Additionally, some schemata have shown poor applicability for manual annotation, with annotation campaigns reporting low inter-annotator agreement (e.g., those by Egbert et al. (2015) and Sharoff (2018)). In addition, when developing a high-coverage schema for web text classification, it is important to acknowledge that writers on the web are not constrained to following genre norms, meaning that some texts will inevitably fall outside any predefined genre class (Sharoff, 2021). While the existing high-coverage schemata rely on a closed set of predefined labels, it would be advisable to include a category *Other*, as suggested by Asheghi et al. (2016).

2.2 Development of Manually-Annotated Genre Datasets

Since the genre classifier is trained on a manually-annotated dataset, the quality of this dataset is critical for model performance. Most importantly, the dataset should be robust – which in this context means that its manual annotation must be reliable and that it should include a diverse set of texts that represent well the diversity of web content. This enables the classifier to generalize beyond the training data to other datasets. There are two stages of the development of training genre datasets that importantly impact the quality of the resulting datasets: the development of the collection of representative texts to be annotated, and a manual annotation campaign.

Based on the collection method, text collections used as the basis for manually-annotated genre datasets can be categorized into designed and crawled corpora (Kilgarriff,

2012). Designed corpora, such as those by Berninger et al. (2008), Boese (2005), M. Santini (2007), and Stubbe and Ringlstetter (2007), are based on manual selection of instances. As a result, they are not representative of real web data, as noted by Sharoff (2010), and the high classification performance on such datasets does not guarantee that the model will generalize well to real web text collections. Crawled corpora, such as those by Laippala et al. (2020) and Sharoff (2018), extract random instances from massive web corpora and represent a "(more or less faithful) snapshot of the web" (Asheghi et al., 2016). Random selection from a large web text collection ensures that instances originate from a wide range of sources and helps prevent bias toward specific topics, authors, or web domains. A further step in this direction was taken by Asheghi et al. (2016), who also ensured temporal diversity of resources by selecting sources published at different time points, preventing that some genres, such as *News*, would be represented solely by instances covering the same topics.

The second key factor in producing a high-quality training dataset is reliable manual annotation. For this task, previous studies employed either a small group of experts (e.g., annotation campaigns in S. M. Santini (2006), Vidulin et al. (2007), and Zu Eissen and Stein (2004)), or crowd-sourcing (e.g., see Egbert et al. (2015)). The advantage of expert annotation lies in the higher reliability of annotators and greater control over the annotation process, whereas crowd-sourcing allows for faster and more cost-effective annotation, enabling creation of larger manually-annotated datasets.

Annotation reliability is typically assessed through inter-annotator agreement, which measures the agreement of two or more annotators on the label of the same instance. However, prior research has shown that achieving high inter-annotator agreement on this task is difficult, and studies rarely achieved an acceptable level of inter-annotator agreement (see for instance Egbert et al. (2015) and Suchomel (2020)), which is defined in our work as Krippendorff's alpha above 0.667, following Krippendorff (2018). These findings highlight the need to improve the annotation methodology to achieve a more reliable annotation.

Table 2.2 presents the most relevant manually-annotated genre datasets developed in previous work. The comparison shows a considerable variation among the datasets based on their schemata, namely, the label granularity varies significantly, with the number of labels ranging from 7 to 292. This further highlights the limited comparability between datasets. A further issue is that many previously developed genre datasets have become difficult to access over time and are no longer available, for example, those by Berninger et al. (2008), M. Santini (2010), Sharoff (2010), and Stubbe and Ringlstetter (2007). Even the datasets that are currently available are not published in certified repositories which would ensure reliable long-term archiving.

More information on the datasets is provided in the journal paper by Kuzman and Ljubešić (2023) that is enclosed in Section 2.4. In this thesis, we focus mainly on high-coverage schemata that could be used to develop a high-coverage genre classifier for web texts. Two families of manually-annotated genre datasets that use a high-coverage schema are openly available and could thus be used in our work, namely the CORE dataset (Egbert et al., 2015) in English and the related datasets in Swedish, French and Finnish (Laippala et al., 2019; Laippala et al., 2020; Repo et al., 2021; Skantsi & Laippala, 2023); and the FTD dataset (Sharoff, 2018) in Russian and English.

The *CORE* dataset (Egbert et al., 2015) is a crawled dataset that consists of 48,420 texts from the Corpus of Global Web-based English (GloWbE) (Davies & Fuchs, 2015). It uses the CORE schema – a hierarchical schema comprising 8 higher-level categories, which cover the situational characteristics or functions of texts, and 47 specific subcategories. The dataset was annotated in a crowd-sourcing campaign with 908 participants, where each instance was annotated by 4 annotators.

The *Functional Text Dimensions* (FTD) dataset (Sharoff, 2018) consists of 1,562 English texts and 1,930 Russian texts from multiple sources: the Pentaglossal corpus (5g) (Forsyth & Sharoff, 2014), which comprises Russian and English texts of known origin (fiction, corporate communication, political debates, TED talks, UN reports, etc.), and random samples of texts from the ukWac web corpus (Baroni et al., 2009) and the Russian GICR web corpus (Piperski et al., 2013). The dataset uses the Functional Text Dimensions (FTDs) schema that consists of 18 broad functional categories. The manual annotation followed a multi-label approach, where annotators rated the presence of each category in the text on a scale from 0 to 2.

The CORE and FTD schemata and datasets cover English as well as several other languages, including Russian, Swedish, French and Finnish. However, no genre-annotated dataset currently exists for Slovenian, which is one of the focal points of our work. Thus, to enable the study of automatic genre identification in Slovenian, a new manually-annotated genre dataset in Slovenian has to be created.

While the CORE and FTD datasets can be useful for developing genre classifiers for English and other supported languages, following their schemata to develop a manually-annotated genre dataset for Slovenian would be challenging. Previous annotation campaigns applying the FTD schema to a new dataset concluded that a high agreement in this task is "hardly possible" (Suchomel, 2020), and an analysis of inter-annotator agreement achieved when annotating the CORE dataset reported that there was no majority agreement between the four annotators on one-third of texts on the level of main categories, and on half of the texts on the level of subcategories (Egbert et al., 2015). When the CORE schema was applied to another dataset, the inter-annotator agreement reached a nominal Krippendorff's alpha of 0.66 and 0.53 for main categories and subcategories, respectively (Sharoff, 2018), which are on the verge and below the threshold of 0.667 for reliable manual annotation.

Table 2.2: Comparison of the most relevant genre-annotated datasets developed in previous work, in terms of their size (number of texts), language (ISO codes), number of classes in the genre schema and multi-label approach. If the schema is hierarchical, the number of classes includes the main classes and sub-classes, e.g., 7/32 stands for 7 main classes and 32 sub-classes. Links for publicly available datasets are provided in the footnotes.

Dataset	Size (texts), Language	Classes	Multi-label
High-coverage datasets			
Hierarchical genre collection (HGC) (Stubbe & Ringlstetter, 2007)	1,280 (EN)	7/32	no
KRYS I (Berninger et al., 2008)	6,300 (EN)	10/70	no
I-EN-Sample, I-RU-Sample (Sharoff, 2010)	250 (EN), 250 (RU)	7	no
SANTINIS-ML (M. Santini, 2010)	2,480 (EN)	15	yes
CORE (Egbert et al., 2015)[1]	48,420 (EN)	8/47	yes
Functional Text Dimensions (FTD) (Sharoff, 2018)[2]	1,562 (EN), 1,930 (RU)	18	yes
FinCORE (Laippala et al., 2019)[3]	2,237 (FI)	8/22[4]	yes
FreCORE, SweCORE (Laippala et al., 2020)[5]	688 (FR), 1,085 (SV)	8/22	yes
Suchomel (2020)	1,974 (EN)	9	yes
LiveJournal dataset (Lepekhin & Sharoff, 2022)	13,629 (RU)	10	no
FinCORE (extended) (Skantsi & Laippala, 2023)[6]	10,754 (FI)	9/30	yes
Information-retrieval datasets			
Dewe et al. (1998)	1,358 (EN)	2/11	no
Roussinov et al. (2001)	1,076 (EN)	116	no
Genre-KI-04 (Zu Eissen & Stein, 2004)[7]	1,239 (EN)	8	no
Lim et al. (2005)	1,328 (KO)	2/16	no
20-Genre Collection (MGC) (Vidulin et al., 2007)[8]	1,539 (EN)	20	yes
7-Web Genre Collection (M. Santini, 2007)	1,400 (EN)	7	no
Other genre datasets			
Boese (2005)	343 (EN)	10	no
Leeds Web Genre Corpus (Asheghi et al., 2016)	4,964 (EN)	20	no

[1]The CORE corpus is available here: https://github.com/TurkuNLP/CORE-corpus and can be queried here: https://www.english-corpora.org/core/.

[2]The FTD dataset is available here: https://github.com/ssharoff/genre-keras.

[3]The FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE.

[4]Based on the annotation guidelines published at https://turkunlp.org/register-annotation-docs.

[5]The FreCORE and SweCORE datasets are available here: https://github.com/TurkuNLP/Multilingual-register-corpora.

[6]The new extended FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE_full

[7]The Genre-KI-04 dataset is available here: https://webis.de/data/genre-ki-04.html.

[8]The 20-Genre Collection can be accessed through the Internet Archive Wayback Machine: https://web.archive.org/web/20120512021524/http://dis.ijs.si/mitjal/genre/

2.3 Machine Learning Approaches for Automating Genre Identification

Prior to the emergence of deep neural classifiers, support vector machines (SVMs) were the most commonly used approach in previous work (Laippala et al., 2021; Laippala et al., 2017; Petrenz & Webber, 2011; Pritsos & Stamatatos, 2018; Rezapour Asheghi, 2015; Sharoff et al., 2010). However, this machine learning method has shown limited generalization capabilities. Sharoff et al. (2010) evaluated the performance of linear SVMs in a cross-dataset classification setting, where the model was trained on one genre dataset and tested on another. The results indicated strong dataset dependence – while the models performed well in the in-dataset scenario, their accuracy on other datasets was mostly below 40%. This does not satisfy our criteria for model robustness which we define as the ability to generalize across datasets and, ideally, across languages. These findings also highlight the importance of evaluating classifiers on multiple datasets, as in-dataset results may not reflect the model's effectiveness in real-world scenarios.

In recent years, SVMs and other non-neural methods have been largely replaced by the Transformer-based pre-trained BERT-like language models which have demonstrated significantly stronger capabilities. Repo et al. (2021) and Rönnqvist et al. (2021) showed that these models can achieve strong performance when trained on smaller training datasets, and even across languages in zero-shot cross-lingual settings. Experiments with varying sizes of the CORE dataset (Egbert et al., 2015) revealed that the performance of a fine-tuned Transformer model plateaued after using approximately 10,000 training instances, indicating that manual annotation of substantially larger datasets is unnecessary (Laippala et al., 2022). Moreover, when the model was evaluated on longer texts, the beginning of the document, that is, the first 512 tokens, was shown to be the most informative part for genre prediction, which dismisses the need for using models designed to deal with longer texts, such as the Longformer language model (Beltagy et al., 2020).

Table 2.3 provides an overview of recent machine learning experiments on datasets from the CORE (Egbert et al., 2015; Laippala et al., 2019; Laippala et al., 2020) and FTD (Sharoff, 2018) genre dataset families, as reported in previous work (Laippala et al., 2022; Repo et al., 2021; Sharoff, 2021). The reported results are hardly comparable, as the experiments were applied on different datasets and used different schemata, machine learning models, and evaluation metrics. Nevertheless, the overview still provides an insight into the state-of-the-art performance of genre classifiers.

Multiple studies have shown the massively multilingual XLM-RoBERTa model (Conneau et al., 2020) to be the most appropriate for this task, achieving micro-F1 scores between 0.72 and 0.84 in the in-dataset classification setting (Repo et al., 2021; Rönnqvist et al., 2021). Recent studies advise that the model is fine-tuned on a combination of multiple genre datasets, which improves its performance and its robustness against topic bias in training data (Lepekhin & Sharoff, 2022; Rönnqvist et al., 2021).

The fine-tuned XLM-RoBERTa model has also shown promising zero-shot cross-lingual performance with F1 scores between 0.61 and 0.69 when trained on English CORE dataset and evaluated on Finnish, French, and Swedish (Repo et al., 2021). The XLM-RoBERTa model trained on an extended multilingual dataset using a modified set of CORE labels was further evaluated in a zero-shot cross-lingual scenario on 11 languages (Henriksson et al., 2024), achieving an average micro-F1 score of 0.66 and an average macro-F1 score of 0.62. These results indicate promising possibilities for the development and application of genre classifiers on less-resourced languages based on multilingual Transformer-based BERT-like language models.

Table 2.3: Overview of the model (in-dataset) performance on the most relevant high-coverage genre datasets, as reported in previous work.

Research	Dataset	Number of labels	Model	Best score (per dataset)
Sharoff (2021)	FTD (English), FTD (Russian) (Sharoff, 2018)	10 (subset of FTD labels)	Bi-directional Long Short-Term Memory (BiLSTM) classifier (Yogatama et al., 2017)	Precision: 0.77, 0.75
Repo et al. (2021)	FinCORE, FreCORE, SweCORE, CORE (Egbert et al., 2015)	7 (subset of CORE main labels)	XLM-RoBERTa large (Transformer model) (Conneau et al., 2020)	F1: 0.73, 0.77, 0.83, 0.76
Laippala et al. (2022)	CORE (Egbert et al., 2015)	56 (CORE main and sub-categories; multilabel approach)	BERT large (Transformer model) (Kenton & Toutanova, 2019)	Micro-F1: 0.68

2.4 Related Paper: Automatic Genre Identification: A Survey

Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09695-8

SURVEY

Automatic genre identification: a survey

Taja Kuzman[1,2] · **Nikola Ljubešić**[1,3]

Accepted: 18 September 2023

Abstract

Automatic genre identification (AGI) is a text classification task focused on genres, i.e., text categories defined by the author's purpose, common function of the text, and the text's conventional form. Obtaining genre information has been shown to be beneficial for a wide range of disciplines, including linguistics, corpus linguistics, computational linguistics, natural language processing, information retrieval and information security. Consequently, in the past 20 years, numerous researchers have collected genre datasets with the aim to develop an efficient genre classifier. However, their approaches to the definition of genre schemata, data collection and manual annotation vary substantially, resulting in significantly different datasets. As most AGI experiments are dataset-dependent, a sufficient understanding of the differences between the available genre datasets is of great importance for the researchers venturing into this area. In this paper, we present a detailed overview of different approaches to each of the steps of the AGI task, from the definition of the genre concept and the genre schema, to the dataset collection and annotation methods, and, finally, to machine learning strategies. Special focus is dedicated to the description of the most relevant genre schemata and datasets, and details on the availability of all of the datasets are provided. In addition, the paper presents the recent advances in machine learning approaches to automatic genre identification, and concludes with proposing the directions towards developing a stable multilingual genre classifier.

Keywords Text genre · Web genre · Automatic genre identification · Genre schemata · Genre datasets · Survey paper

✉ Taja Kuzman
taja.kuzman@ijs.si

Nikola Ljubešić
nikola.ljubesic@ijs.si

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

[3] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

Published online: 16 November 2023

1 Introduction

Genres are text categories, defined by the author's purpose, common function of the text, and the text's conventional form (Orlikowski & Yates, 1994). They have been studied already in Antiquity, as the Greek philosophers and orators recognized that the manner in which the information is conveyed plays an important role on the receivers' understanding of the message, as well as the efficiency of communication (Kwasnik & Crowston, 2005). As genre involves multiple discourse, language and textual phenomena, it has been studied by researchers from various fields. The term is widely used in linguistics, rhetoric, literary theory, discourse studies, media theory, as well as computational linguistics and information retrieval studies (Chandler, 1997). The importance of providing information on genres included in a text collection was also recognized by the authors of reference corpora, such as the British National Corpus (BNC) (Davies, 2004) and the Corpus of Contemporary American English (COCA) (Davies, 2008).

Since the advent of the World Wide Web, there has been an increased interest in identifying genres in text collections, as the web allowed querying and collecting unprecedented quantities of texts. There has been large interest in genre identification in the area of information retrieval (see Stubbe and Ringlstetter 2007; Eissen and Stein 2004; Vidulin et al. 2007; Roussinov et al. 2001; Finn and Kushmerick 2006; Boese 2005; Priyatam et al. 2013; Stein et al. 2010), as the information on the genre in a query in information retrieval tools would allow users to find the texts that are relevant for them more efficiently (Crowston et al., 2010). Furthermore, collecting texts from the web has allowed a significantly faster and cheaper creation of corpora which are essential for language research and development of advanced language resources and technologies, especially in cases of under-resourced languages. However, in contrast to corpora that were carefully manually selected, web corpora are built in an automated way due to which their composition remains unknown (Baroni et al., 2009).

To obtain the information on the genre of specific texts in large collections of text, automatic genre identification (AGI), a text classification method which categorizes texts into genres, is the only feasible option due to the high data volume. AGI is primarily researched in the field of language technology which is concerned with computational processing of human languages and which encompasses computational linguistics and natural language processing (NLP), combining linguistics and computer science, especially artificial intelligence. As the authors of smaller, manually collected collections of texts are usually aware of the source of texts and their types, automatic genre identification is mostly studied to be applied on large web-based text collections where the source of texts is unknown. That is also why most of genre-annotated datasets surveyed in this paper contain texts from the web. However, since many high-coverage genre schemata, presented in this work, aim to cover all of the genre categories found in texts, the results of web genre identification research can be applied to collections of texts which do not originate from the web as well.

In addition to the benefits of genre identification for information retrieval and corpus creation and curation, it was shown that annotating documents with genre is also beneficial in many other NLP tasks. For instance, in part-of-speech (PoS) tagging of web corpora, Giesbrecht and Evert (2009) showed that using different granularity of PoS sets depending on the genres improves the tagging accuracy. Stewart and Callan (2009) showed that automatic genre identification improves automatic summarization, as the genre-oriented and goal-focused summarization algorithms outperformed other summarization methods. In addition to this, genres were revealed to be beneficial for machine translation (Van der Wees et al., 2018) and zero-shot dependency parsing (Müller-Eberstein et al., 2021). Recently, genre has also been studied in the area of information security, where it can be used for automated genre-aware assessment of credibility of web documents (Agrawal et al., 2019).

Obtaining information on genres can be of great value in a wide range of disciplines, including linguistics, corpus linguistics, computational linguistics, natural language processing, information retrieval and information security. Therefore, automatic genre identification models would be a great resource, as manual annotation of thousands of texts with genre labels is expensive and time-consuming, and, despite the best efforts, it can often result in unreliable annotations. However, although various attempts have been made to provide schemata, genre-annotated datasets and classifiers, there are no reference datasets (Sharoff, 2010) that can be directly used for building an AGI classifier which could be applied to any collection of web documents. In previous work, researchers used various genre definitions, annotation practices and corpora collection methods, each of them having its advantages and disadvantages. Thus, when approaching the automatic genre identification (AGI) task, one must not only ponder the appropriate machine learning strategy, but needs to carefully define each step of the data collection, schema creation and data annotation process.

The objective of this paper is to present a detailed survey of the existing research on automatic genre identification and available genre datasets. The rest of the paper is organised as follows: Sect. 2 presents various terms and definitions used to describe genres, which are the basis of genre schemata, discussed in Sect. 3. Section 4 compares existing genre-annotated datasets given the data collection methods and annotation approaches. The most frequently used datasets are presented in more detail. Section 5 presents an overview of text classification machine learning methods typically used for the automatic genre identification task, the results of various experiments, and the main challenges for automatic genre identification. Finally, Sect. 6 concludes the paper by summarising its findings, and proposing directions for improvements in the area.

2 Genre definition

One of the main challenges of studying genre is that there is no consensus among researchers on the definition of the genre phenomenon, or how to differentiate between genres, what qualifies for a genre class, and, finally, what the optimal genre schema is. Even more, there is no consensus on the term used for this phenomenon.

While many researchers use the term "genre" and describe this task as "automatic genre identification" (AGI) or "web genre identification" (WGI), others avoid this term and prefer derived analogues, such as "registers" (see Egbert et al. 2015 and further research by Laippala et al. 2019; Repo et al. 2021; Ronnqvist et al. 2021), "functional text dimensions" (see Sharoff 2018) or "text types" (see Stubbs 1996). While extensive work has been dedicated to navigate this "terminological maze", as Moessner (2001) put it, especially to distinguish between registers and genres (see Biber and Conrad 2019, Sharoff 2021, Lee 2002A), some researchers use the two terms interchangeably and argue that the distinction between them is "not relevant, and the choice between genre and register comes down to personal preference or tradition" (Egbert et al., 2015). However, in linguistics and especially sociolinguistics the term "register" is used to describe a different aspect of language variation, connected with conversational context. Frequent register categories are "academic language", "non-standard language", "formal language" and so on (see Lukin et al. 2011). According to Lee (2002A), the largest difference between the concepts of genres and registers is that genres describe whole texts, while registers are defined based on internal linguistic patterns, which are independent of text-level structures.

In addition to this, another challenge is how to define genres, an endeavour described by Chandler (1997) as a "theoretical minefield". In general, most definitions describe genre classes based on the socially recognized form of a document and its intended communicative purpose (Kwasnik & Crowston, 2005). In addition to this, some definitions also consider the target audience (Eissen & Stein, 2004), expectations of the reader (Santini, 2006), style (Finn & Kushmerick, 2006; Argamon et al., 1998), or the content of the texts (Rosso, 2008). To identify genres, researchers observe their intrinsic features, i.e., the linguistic and other "look'n'feel" features in the text (text-internal perspective), their extrinsic features, that is the function of the texts (text-external or functional perspective), or both (see Sharoff (2010, 2021) for a detailed discussion on both perspectives).

Moreover, studying genres, especially genres of digital texts, is a "formidable undertaking", to cite Kwasnik and Crowston (2005), due to fundamental difficulties connected with the genre notion itself. Firstly, conventional characteristics of a genre category can change over time and instances can be more or less prototypical in a specific time period (see Santini et al. (2010), Santini (2006)). The evolution of genres was especially notable after the introduction of the World Wide Web, when researchers observed emergence of new genres, unique to the web, and existing genres that evolved to adapt to the new medium (Williams & Crowston, 2000). Although many genres of web documents exist also in traditional texts, genres on the web are less fixed (Kwasnik & Crowston, 2005). Secondly, some web documents can be hybrids, which means that a single text can display characteristics of multiple genres, e.g., an advertorial, a document that has the form of a news article, but the purpose of a promotion of a product (see Sharoff (2021) and Repo et al. (2021)). Thirdly, some web documents might not have any discernible characteristics of a genre class. They can be too short to reveal the purpose of the communication. They can also be of an emerging genre, or a specialized genre with which the reader is not familiar with, e.g., a shipping invoice, used by accountants (see Williams and Crowston (2000)). In addition to this, it is argued that genre can only be realized

in a complete text (Biber & Conrad, 2019), which is not always available after the extraction of the text from a web page, a process that often includes removal of boilerplate content and other imperfect processing methods (see Pomikálek (2011)).

3 Genre schemata

Eissen and Stein (2004) noted that an "inherent problem of web genre classification is that even humans are not able to consistently specify the genre of a given page". Text instances can be more or less prototypical examples of its genre classes, can be hybrids or can lack characteristics of any genre. Thus, it is crucial for the reliability of a genre-annotated dataset that the genre schema supports the annotators so that their decisions are as consistent as possible. As no reference schema exists, most studies use their own, devised in accordance with the aim of the research. Schemata, used in previous work, vary significantly in terms of the number of classes, which range from seven (Santini, 2007; Sharoff, 2010; Lee & Myaeng, 2002b) to more than a hundred (Roussinov et al., 2001) or almost 300 classes (Crowston et al., 2010). They are either hierarchical (Stubbe & Ringlstetter, 2007; Egbert et al., 2015) or not (Asheghi et al., 2016; Sharoff, 2018).

Most existing schemata consist of around 10 to 20 categories, while some of them comprise of more than 50 genre classes (see Crowston et al. (2010); Roussinov et al. (2001); Williams and Crowston (2000); Egbert et al. (2015); Berninger et al. (2008)). Larger granularity usually increases the annotation complexity which results in a lower inter-annotator agreement and a less reliable dataset (see Sharoff (2010)). However, it should be noted that although a smaller set results in broader categories, they should still be limited by functional and form constraints so that the texts belonging to them share certain distinguishing properties, as argued by Asheghi et al. (2016). As in the case of a high granularity, too broad categories can also negatively impact the reliability of the annotated dataset and can be less useful for automatic identification. In addition to this, if the aim of a genre schema is to capture the entire diversity of genres on the web, Sharoff (2021) warned that genres are not fixed and writers are not constrained to following genre norms, thus, one should be aware that there will always exist web documents which do not fit under any genre class. While most studies used a pre-defined closed set of labels, some previous work took this issue into consideration, i.e., Asheghi et al. (2016) and Kuzman et al. (2022b) included a category *Other* in the schema, and Pritsos and Stamatatos (2018) used an open-set classification approach, allowing for a document not to be associated with any genre class.

Most of the schemata were developed following the top-down approach, which is based on theoretical principles, researchers' knowledge and understanding of the domain, and categorization from previous research. In regard to this approach, Rehm et al. (2008) proposed that the researchers should take advantage of accumulated expert knowledge, to avoid choosing the genre set based solely on how they perceive the classes, because the end users without a linguistic education might understand genre categories differently than genre researchers. By deriving the schema from category sets proposed by various research groups, this approach leads

to an improved set of categories that have already been proven to be applicable to the variety of language in actual document collections. An alternative approach is the bottom-up approach, taken up by some researchers (Crowston et al., 2010; Dewe et al., 1998; Eissen & Stein, 2004; Egbert et al., 2015), which is based on surveys of web users. This approach revealed to be problematic as the respondents were not consistent, they had difficulties with naming genres or used general or unspecific terms (see Crowston et al. (2010)).

The main factor defining the composition of genre schemata in previous works was the aim with which the genre dataset was constructed. Some researchers aimed to create a schema which would cover all possible genre variation found in (mostly web) texts. Such datasets allow for creation of general-purpose classifiers which would be able to classify any text into some genre class. In contrast, other researchers were not interested in covering all of the diversity found in text collections. They rather focused on a smaller set of specific genres that they deemed to be the most useful for search engine users who would find texts more efficiently based on genre criteria. In addition, some other schemata and corresponding genre datasets were developed with the aim to study characteristics of specific genres, such as poetry and prose, or with the aim of improving manual annotation. Thus, based on the aim of genre research, we divided the genre schemata into the following groups:

1. High-coverage schemata (complete genre taxonomy) which aim to cover the entire genre composition of the web. These studies mostly used larger sets of specific classes (see Stubbe and Ringlstetter (2007); Sharoff (2018); Egbert et al. (2015); Laippala et al. (2019); Kuzman et al. (2022b); Suchomel (2020)), a smaller set of broader categories (see Sharoff (2010)) or a combination of the two (see Santini (2010)).
2. Information retrieval schemata, created with the aim to provide genre identification inside the search engines. Thus, they consist of genre categories that are deemed to be useful for the search engine users (see Eissen and Stein (2004); Vidulin et al. (2007); Roussinov et al. (2001); Dewe et al. (1998); Lim et al. (2005); Santini (2007)).
3. Schemata, developed with other aims: limited sets of categories used in studies which analyse one or a few genre classes, e.g., a smaller set of genres of interest (Boese (2005); Lee and Myaeng (2004); Asheghi et al. (2016)), academic genres (Rehm, 2002), e-shop genres (Levering et al., 2008), poetry and prose genres (Shavrina, 2019), news articles and reviews (Finn & Kushmerick, 2006), and home pages (Kennedy & Shepherd, 2005).

The most relevant schemata are presented in more details in Sect. 4.3, together with the datasets for which they were used.

To explore which categories appear most frequently in the schemata, we analysed 16 of the previously mentioned schemata, i.e., schemata from the works of Egbert et al. (2015), Asheghi et al. (2016), Stubbe and Ringlstetter (2007), Eissen and Stein (2004), Santini (2007), Vidulin et al. (2007), Sharoff (2010), Laippala et al. (2019), Santini (2010), Sharoff (2018), Dewe et al. (1998), Lim et al.

Table 1 Most frequent genre labels from 16 genre schemata

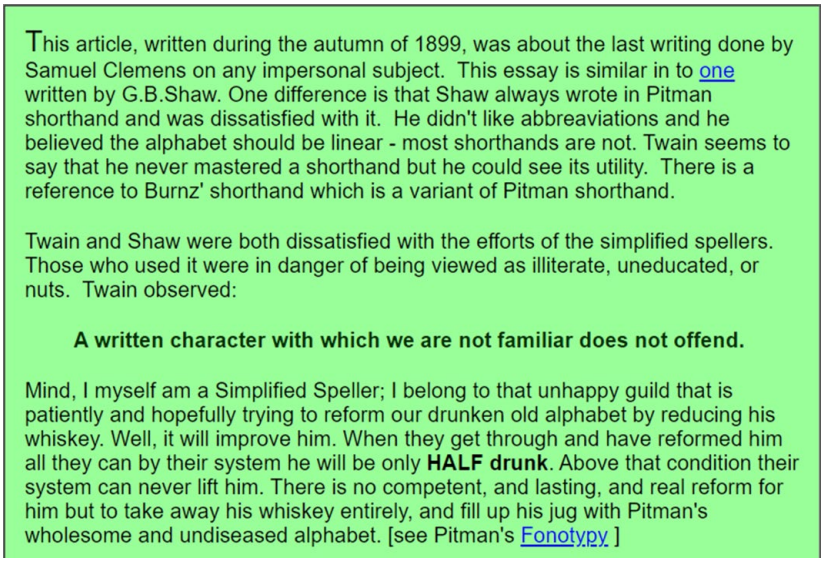
Genre label	Occurrences in schemata
FAQ	9
Instruction	7
News	6
Legal	6
Personal home page	6
Review	6
Information	5
Discussion	5
Research article	5
Interview	5
Lyrical	5
Poem	4
Recipe	4
E-shop	4
Editorial	4
Promotion	3
Link Collection	3
Journalistic	3
Blog	3
Prose	3

Labels with very similar names, e.g., "FAQ" and "FAQs" or "news" and "news/reporting", were counted as the same label name

(2005), Boese (2005), Lee and Myaeng (2002b), Suchomel (2020), Kuzman et al. (2022b). Together, the schemata have 280 genre labels, out of which 214 have unique names. After merging genre labels that have very similar names, such as "FAQ" and "FAQs" or "news" and "news/reporting", there are 177 unique genre labels, out of which 129 appear only in one schema and 28 only in two schemata. Table 1 shows the 20 labels that are used in at least 3 schemata. Out of these labels, the most frequent categories are *FAQ* and *Instruction*, appearing in 7 or more schemata, and *News*, *Legal*, *Personal Home Page* and *Review*, appearing in 6 schemata.

Schemata also vary in terms of the simplicity of names, used for genre labels. As shown in Table 1, the most frequent categories are named with common words with the aim of making the labels understandable to the end users, e.g. *Instruction*, *News* or *Interview*. In contrast, some researchers opted for more abstract label names, such as *content delivery* (Vidulin et al., 2007), *resources* (Stubbe & Ringlstetter, 2007), *recreation* (Sharoff, 2010), *commpuff* (Sharoff, 2018).

While Table 1 shows which category names are most frequently used in genre schemata, we should note that this is a harsh generalization of the existing categories, primarily for purposes of obtaining an overview. The reality in the


This article, written during the autumn of 1899, was about the last writing done by Samuel Clemens on any impersonal subject. This essay is similar in to one written by G.B.Shaw. One difference is that Shaw always wrote in Pitman shorthand and was dissatisfied with it. He didn't like abbreaviations and he believed the alphabet should be linear - most shorthands are not. Twain seems to say that he never mastered a shorthand but he could see its utility. There is a reference to Burnz' shorthand which is a variant of Pitman shorthand.

Twain and Shaw were both dissatisfied with the efforts of the simplified spellers. Those who used it were in danger of being viewed as illiterate, uneducated, or nuts. Twain observed:

A written character with which we are not familiar does not offend.

Mind, I myself am a Simplified Speller; I belong to that unhappy guild that is patiently and hopefully trying to reform our drunken old alphabet by reducing his whiskey. Well, it will improve him. When they get through and have reformed him all they can by their system he will be only **HALF drunk**. Above that condition their system can never lift him. There is no competent, and lasting, and real reform for him but to take away his whiskey entirely, and fill up his jug with Pitman's wholesome and undiseased alphabet. [see Pitman's Fonotypy]

Fig. 1 An example of a text from the category *Article* in the Genre-KI-04 (Eissen & Stein, 2004) dataset

underlying datasets is much more complicated. Labels that have same or similar names do not necessarily cover similar texts, as this depends on the researchers' and annotators' understanding of the labels. Moreover, the texts inside similarly named classes in different datasets could differ significantly also based on the approaches, with which the datasets were collected, as described in the following section.

The case of journalistic genres illustrates well the differences between the schemata and shows why it is infeasible to merge any datasets, even if their labels have similar names. For instance, the Genre-KI-04 dataset (Eissen & Stein, 2004) includes a label *Article*, for which one might assume that it represents news articles. However, in the documentation accompanying the dataset, the label is defined as "Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters". The label thus includes a large variety of texts, including opinionated texts, such as the instance in Fig. 1. In the 20-Genre Collection (MGC) (Vidulin et al., 2007), a genre that could be connected to news articles based on the name is *Journalistic*. Its definition further confirms this: "conveys mostly objective information on current events" (Vidulin et al., 2007). However, on their website,[1] authors included the following types of texts under the *Journalistic* genre: news, reportages, editorials, interviews and reviews, indicating that the class is quite broad and includes objective as well as subjective texts. The great inner variation between the texts, included in this genre, can be seen in Fig. 2. In the CORE dataset (Egbert et al., 2015), the news article genre is divided into two categories – *News Report/Blog* and *Sports Report* – based on the topic of articles. In addition,

[1] The website can be accessed through the Internet Archive Wayback Machine: https://web.archive.org/web/20120513113203/http://dis.ijs.si/mitjal/genre/list_of_genres.htm

Fig. 2 Two instances, labelled with the category *Journalistic* in the 20-Genre Collection (MGC) dataset (Vidulin et al., 2007)

A message to all the Olympic moaners... BELT UP! It's a crisis. It's a complete disaster. It's not just a shambles out there, people; it's an Olympic omnishambles. The Games are six days away and nothing works, absolutely everything is broken and the only solution is to cancel the event and send all those arriving at Heathrow straight back home on the next available flight. Why? Because according to reports, any fool venturing into London will die of carbon monoxide poisoning as they sit in month-long traffic jams. Or drown in their body sweat on overcrowded Tube trains. Or sink into oblivion trying to negotiate the mud flats otherwise known as the Olympic Park. The mobile phone networks will fail, the internet will collapse into a black hole in cyberspace, pickpockets will steal everything, including your kidneys, and the entire country will end up bankrupt. It's a nightmare - and all because of the 2012 Games. You have been warned! Stop moaning! Great

Ash tree disease found for first time in Wales Forestry Commission Wales has confirmed the discovery of a serious disease of ash trees for the first time in Wales. Chalara dieback of ash has been found in a small, privately-owned young broadleaved woodland in Carmarthenshire. The trees, which were planted in 2009, are still quite small and measures are being put in place to minimise the risk of the disease spreading to the wider environment. A containment notice will be issued shortly for the site and FC Wales will work with the woodland owner to deal with the infected trees. The discovery follows a rapid survey of the whole of Wales over the past four days to check the condition of the country's ash trees as plant health authorities across the United Kingdom stepped up their efforts to tackle the disease, caused by the fungus Chalara fraxinea. FC Wales and Welsh

Fig. 3 Two instances, labelled with the category *News Report/Blog* in the CORE dataset (Egbert et al., 2015)

the schema also includes a class, named *Magazine Article*. The *News Report/Blog* genre is defined in the annotation guidelines[2] as a "news report written by journalists and published by news outlets" with the purpose to report on recent events. Despite a clear description, instances in this class vary from objective reports on recent events to opinionated blog-like posts, as can be seen in Fig. 3.

The examples in Figures show that a detailed exploration of texts inside the collections would be necessary to be able to compare the genre schemata thoroughly, which is outside of the scope of the paper. What is more, more detailed analyses of genre classes and the compatibility of datasets are hampered by unavailability of the datasets that use the analysed schemata.

[2] https://turkunlp.org/register-annotation-docs/register/NA-ne.html

4 Genre-annotated datasets

This section discusses the process of constructing datasets that are manually annotated for genre. They are the basis for training machine learning models for automatic genre identification. Some genre researchers used traditional general-purpose corpora that have genre information, such as Corpus of Contemporary American English (COCA) (Davies, 2008) (see Ströbel et al. (2018)), British National Corpus (BNC) (Lee, 2002A) (see Sharoff (2010)) and Brown Corpus (Kučera & Francis, 1967) (see Karlgren and Cutting (1994)). However, these datasets do not represent well the challenges for automatic genre identification imposed by less curated web text collections, where there is much wider variation between styles and quality of texts (Sharoff, 2010). That is why in the great majority of studies, the genre researchers decided to construct their own datasets based on web data. The process of building genre-annotated datasets regularly starts with a selection of data to be manually annotated, as described in Sect. 4.1. Next, manual annotation for genre is performed, as discussed in Sect. 4.2. We wrap up this section with an overview of various types of manually annotated datasets in Sect. 4.3.

4.1 Data Selection

Most datasets that are manually annotated for genre consist of texts originating from the web. The texts are either taken from existing web corpora, or new document collections are constructed in a manual or an automated manner.

As genre-annotated datasets were created with different aims, and financial and technical constraints, they largely differ in terms of size, collection methods and document format. In many studies, the datasets were limited to document formats which are easier to obtain from the web and to process with automated methods, such as HTML. This influences the genre distribution in the final dataset, as by not including PDF files or Word documents, certain genres for which the preferred format is PDF, such as research papers and catalogues, are likely under-represented (Santini et al., 2010).

Based on the collection method, web corpora can be categorized into designed and crawled corpora (Kilgarriff, 2012). In designed corpora, texts are collected based on certain criteria, defined by the researchers. Although this collection method facilitates the annotation and machine classification, Sharoff (2010) argues that designed corpora are not representative of the actual genre distribution on the web. They are thus less useful for training a genre classifier that would be applied on a new text collection. In contrast, crawled corpora represent a "(more or less faithful) snapshot of the web" (Asheghi et al., 2016). Examples of designed corpora are those by Stubbe and Ringlstetter (2007), Santini (2007), Boese (2005), Berninger et al. (2008), while Laippala et al. (2020), Kuzman et al. (2022b) and Sharoff (2018) used samples of crawled corpora. Rehm et al. (2008) proposed using both methods, which was done by Asheghi et al. (2016). They first collected a designed corpus consisting of prototypical instances to test their method of manual annotation, and then extended their research to a random collection of web texts to analyse the coverage of the genre

schema and the reliability of annotation in a more realistic setting. Another commonly used method for data collection is retrieving top hits of automated queries made to a search engine (see Egbert et al. (2015); Sharoff (2010); Lim et al. (2005)). However, one should be aware that the final composition of such collection might be impacted by the choice of keywords in the queries and by the searching and ranking algorithms used by the search engines which favour more popular pages (see Santini et al. (2010); Asheghi et al. (2016)). This can result in a bias towards certain genres, such as e-shops. Furthermore, this method allows collecting only documents that can be found on the searchable web (Egbert et al., 2015).

Being aware of the impact of the collection method on the resulting dataset, Asheghi et al. (2016) and Lim et al. (2005) devoted attention to assuring that the texts from the crawled corpus originate from a wide variety of sources, to avoid biases towards certain topics, authors or sources, thereby guaranteeing generalizability of the resulting datasets. Furthermore, Asheghi et al. (2016) and Kuzman et al. (2022b) assured temporal diversity of sources, by selecting sources published at several time points, as some genres, such as *News*, have strong temporal connection to topics.

4.2 Annotation practices

In addition to the corpus collection methods that the datasets are based on, one of the most crucial factors for producing a high-quality dataset that can be used for machine learning experiments is reliable manual annotation. The annotation can be performed by a smaller group of experts, such as in the case of Santini (2006); Vidulin et al. (2007); Eissen and Stein (2004). However, as the task is very time-consuming, annotation campaigns in previous research mostly resulted in relatively small datasets, consisting of less than 2,000 texts. To achieve faster and cheaper manual annotation, some researchers leveraged the advent of crowd-sourcing. This is how two of the biggest English genre-annotated corpora were created, the Leeds Web Genre Corpus (Asheghi et al., 2016) and the CORE corpus (Egbert et al., 2015). The decision to opt for crowd-sourcing was based on the previous findings (see Snow et al. (2008)) which compared expert and crowd-sourcing annotation on various tasks and concluded that if four non-experts annotate the same instance, the final annotations are often comparable to the expert-annotation quality. In addition to this, Asheghi et al. (2016) argue that this is the most appropriate annotation method, because it avoids potential bias from the experts who both developed the genre schema and annotated the dataset.

Furthermore, annotation strategies vary based on how researchers dealt with hybridity of web texts. As mentioned in the Introduction, classifying web texts in genre categories can be challenging, because some texts, known as hybrid texts, can have characteristics of multiple genres. The hybridity of texts might be intentional, which is the case of an advertorial – a promotion of a product written in a form of a news article to appear more trustworthy to the readers. It might also occur due to authors' lack of expertise or willingness to follow traditionally accepted genre characteristics (Sharoff, 2021). In contrast to non-digital texts which are often reviewed

before being printed and published, a much larger number of texts on the web does not undergo any editorial control which would compel the authors to follow genre norms.

In addition to hybrid texts where characteristics of multiple genres are intertwined, there exist multi-class documents, which consist of a selection of functionally-separated texts of different genres that were published on the same web page and thus belong to a single document (Stubbe & Ringlstetter, 2007). An example of a multi-class document is a document from a web page consisting of a news article, followed by the readers' comments. A special and the most-challenging case of multi-class documents are documents where text in another genre is embedded in another text (see Williams and Crowston (2000)), such as a blog post that includes an entire letter. To assure a reliable annotation, challenges with hybrid texts and multi-class documents need to be addressed in the data selection and annotation procedure. Some previous studies tackled the issue with hybrid texts by allowing multi-label annotation, i.e., assigning multiple labels to a single text (Vidulin et al., 2007; Laippala et al., 2019; Sharoff, 2018; Suchomel, 2020; Egbert et al., 2015; Kuzman et al., 2022b).

The reliability of the annotation procedure is commonly evaluated based on the inter-annotator agreement, i.e., agreement of two or more annotators on the label of the same instance. That is why most of the datasets were annotated by at least 2 annotators. However, their reliability cannot be directly compared, because different methods were used to calculate the agreement: percentage agreement (in Asheghi et al. (2016), Roussinov et al. (2001), Williams and Crowston (2000), Berninger et al. (2008)), Fleiss' kappa (Fleiss, 1971) (in Asheghi et al. (2016), Egbert et al. (2015)), Cohen's kappa coefficient (Cohen, 1960) (in Vidulin et al. (2007)), Krippendorff's alpha (Krippendorff, 2018) (in Vidulin et al. (2007); Sharoff (2018); Suchomel (2020); Kuzman et al. (2022b)), Jaccard similarity coefficient (in Suchomel (2020)), accuracy (in Suchomel (2020)) and F1 score (in Repo et al. (2021)).

Despite the best efforts in creating optimal genre schemata, with annotation guidelines and surveys that provide guidance to the annotators, previous studies revealed that achieving high inter-annotator agreement is very challenging. For instance, Suchomel (2020) concluded that a high agreement in this task is "hardly possible". This is also the case for two of the datasets that have been recently most frequently used, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Functional Text Dimensions (FTD) dataset (Sharoff, 2018). For CORE (Egbert et al., 2015), an analysis of the inter-annotator agreement revealed that there was no majority agreement, i.e., agreement between at least three of four annotators, on one-third of texts on the level of 8 main categories, and on half of the texts on the level of 53 subcategories. When Sharoff (2018) applied the CORE schema and annotation process on another dataset, the inter-annotator agreement reached a nominal Krippendorff's alpha of 0.66 and 0.53 for main categories and subcategories respectively, which is on the verge and below the acceptable threshold of 0.67, defined by Krippendorff (2018). For the FTD dataset, the inter-annotator agreement initially revealed encouraging results, reaching Krippendorff's alpha above 0.76 (Sharoff, 2018). However, when the annotation was extended to another

dataset in subsequent research by Suchomel (2020), the minimal acceptable Krippendorff's alpha was not reached despite using a smaller set of labels. Recently, based on the findings from the previously discussed work, Kuzman et al. (2022b) considered assuring a reliable annotation at every step of their research. They devised their genre schema based on the categories from previous schemata so that the categories would be widely used and the label names would be recognizable to the annotators. The authors also opted for expert annotators instead of crowd-sourcing, and devised a decision tree survey and detailed guidelines with examples which led the annotators through their decision process. The results of the annotation campaign showed that this endeavour resulted in an inter-annotator agreement up to the nominal Krippendorff's alpha of 0.71.

To avoid time-consuming and costly manual annotation, Santini (2006) proposed "annotation by objective sources". This means that she identified genre-specific web domains, such as e-shops, and considered all texts from the domain as automatically belonging to the same genre without any further manual annotation. Automatic annotation was also applied to texts that contained the name of the genre in the title. This approach is based on the idea that if a text originates from a web domain which specializes in publishing texts of one genre, this means that web users collectively consider the text to belong to this genre. This approach was recently taken up by Lepekhin and Sharoff (2021) as well, who named it "natural annotation".

If researchers opted to use an existing genre schema for performing automatic genre identification experiments, they most often chose those from the few openly available datasets. For instance, two earlier schemata, Genre-KI-04 (Eissen & Stein, 2004) and 7-Web-Genre-Collection (Santini, 2006), were often used for evaluation of automatic genre identification algorithms (Jebari, 2021; Pritsos & Stamatatos, 2018; Kanaris & Stamatatos, 2007, 2009; Mason et al., 2009). However, today, most AGI experiments, as well as new genre datasets, are based on the schemata of the Corpus of Online Registers of English (CORE) (Egbert et al., 2015), the Functional Text Dimensions (FTD) approach (Sharoff, 2018), or the GINCO corpus (Kuzman et al., 2022b). The datasets and their schemata are described in more details in the following section.

4.3 Comparison of datasets

An overview of genre-annotated datasets with information on their size, language, number of classes and hierarchy depth, annotation approach and availability is shown in Table 2. The most relevant datasets, divided into high-coverage datasets, information retrieval datasets and other datasets, are described in more details in the following subsections.

As shown in Table 2, less than half of the datasets are available online for use in further experiments. As none of the datasets, except the GINCO dataset (Kuzman et al., 2022b), is published in a repository, they are hard to find and can become unavailable over time. For instance, previously most often used datasets, e.g., the 7-Web Genre Collection (Santini, 2007), the Hierarchical genre collection (Stubbe & Ringlstetter, 2007), KRYS I (Berninger et al., 2008) and others, used to be

Table 2 Overview of the genre-annotated datasets, compared in terms of size (number of texts), language (ISO codes), number of classes in the genre schema and multi-label approach

Dataset	Size (texts), Language	Classes	Multi-label
Dewe et al. (1998)	1358 (EN)	2/11	No
Stamatatos et al. (2000)	250 (EL)	10	No
Roussinov et al. (2001)	1076 (EN)	116	No
Lee and Myaeng (2002b)	7828 (KO), 7615 (EN)	7	Yes
Genre-KI-04 (Eissen & Stein, 2004)[a]	1239 (EN)	8	No
Lim et al. (2005)	1328 (KO)	2/16	No
Boese (2005)	343 (EN)	10	No
Freund et al. (2006)	800 (EN)	16	No
7-Web Genre Collection (Santini, 2007)	1400 (EN)	7	No
20-Genre Collection (MGC) (Vidulin et al., 2007)[b]	1539 (EN)	20	Yes
Hierarchical genre collection (HGC) (Stubbe & Ringlstetter, 2007)	1280 (EN)	7/32	No
KRYS I (Berninger et al., 2008)	6300 (EN)	10/70	No
Maeda and Hayashi (2009)	870 (JA)	7	No
SANTINIS-ML (Santini, 2010)	2480 (EN)	15	Yes
Syracuse dataset (Crowston et al., 2010)[c]	3027 (EN)	292	No
I-EN-Sample, I-RU-Sample (Sharoff, 2010)	250 (EN), 250 (RU)	7	No
CORE (Egbert et al., 2015)[d]	48,420 (EN)	8/47	Yes
Leeds Web Genre Corpus (Asheghi et al., 2016)	4964 (EN)	20	No
Functional Text Dimensions (FTD) (Sharoff, 2018)[e]	1562 (EN), 1,930 (RU)	18	Yes
FinCORE (Laippala et al., 2019)[f]	2237 (FI)	8/22[g]	Yes
FreCORE, SweCORE (Laippala et al., 2020)[h]	688 (FR), 1085 (SV)	8/22	Yes
Suchomel (2020)	1974 (EN)	9	Yes
LiveJournal dataset (Lepekhin & Sharoff, 2022)	13,629 (RU)	10	No
GINCO (Kuzman et al., 2022b)[i]	1002 (SL)	24	Yes
FinCORE (extended) (Skantsi & Laippala, 2023)[j]	10,754 (FI)	9/30	Yes

If the schema is hierarchical, the number of classes includes the main classes and sub-classes, e.g., 7/32 stands for 7 main classes and 32 sub-classes

[a]The Genre-KI-04 dataset is available here: https://webis.de/data/genre-ki-04.html.

[b]The 20-Genre Collection can be accessed through the Internet Archive Wayback Machine: https://web.archive.org/web/20120512021524/http://dis.ijs.si/mitjal/genre/

[c]Details on the dataset are based on the description in Rezapour Asheghi (2015).

[d]The CORE corpus is available here: https://github.com/TurkuNLP/CORE-corpus and can be queried here: https://www.english-corpora.org/core/.

[e]The FTD dataset is available here: https://github.com/ssharoff/genre-keras.

[f]The FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE.

[g]Based on the annotation guidelines published at https://turkunlp.org/register-annotation-docs.

[h]The FreCORE and SweCORE datasets are available here: https://github.com/TurkuNLP/Multilingual-register-corpora.

[i]The GINCO dataset is available here: http://hdl.handle.net/11356/1467.

[j]The new extended FinCORE dataset is available here: https://github.com/TurkuNLP/FinCORE_full

Main Category	Journalism	Literature	Information	Documentation	Dictionary	Communcation [*sic*]	Nothing
Sub-category	Commentary	Poem	Science Report	Law	Person	Mail, Talk	Nothing
	Review	Prose	Explanation	Official Report	Catalog	Forum, Guestbook	
	Portrait	Drama	Receipt	Protocol	Ressources	Blog	
	Marginal Note		FAQ		Timeline	Form	
	Interview		Lexicon, Word List				
	News		Bilingual Dictionary				
	Feature Story		Presentation				
	Reportage		Statistics				
			Code				

Fig. 4 Hierarchical schema, proposed by Stubbe and Ringlstetter (2007) and used in The hierarchical genre collection (HGC)

available at a WebGenreWiki website which was last edited in 2012 and ceased to be available in 2015.[3] Table 2 includes non-English datasets which are in general rare. Although the genre is mainly studied for English, there exist some research groups who work towards contributing datasets in other languages: Sharoff and colleagues from the University of Leeds, UK, with interest in Russian (Sharoff, 2010, 2018, 2021; Lepekhin & Sharoff, 2022) and Arabic (Bulygin & Sharoff, 2018), and the TurkuNLP Group from the University of Turku, Finland, with interest in extending the CORE schema to other languages to leverage the promising advances in cross-lingual learning, providing Finnish (Laippala et al., 2019; Skantsi & Laippala, 2023), Swedish and French (Laippala et al., 2020; Repo et al., 2021) datasets, and smaller evaluation datasets in 8 additional languages (Laippala et al., 2022b). In addition to this, there exist genre-annotated datasets for Korean (Lim et al., 2005; Lee & Myaeng, 2002b), Japanese (Maeda & Hayashi, 2009), Greek (Stamatatos et al., 2000) and Slovene (Kuzman et al., 2022b).

4.3.1 High-coverage datasets

In this section, we present the most relevant high-coverage datasets, i.e., datasets created with the aim of capturing all of the genre variation that can be found in (mostly web) texts. This is reflected in genre schemata, which consist either of a large number of specific categories or a smaller number of broader genre classes.

The hierarchical genre collection (HGC) (Stubbe & Ringlstetter, 2007) is one of the first datasets annotated with a hierarchical schema. The schema was devised following a combination of a top-down and bottom-up approach. It was based on the genre set proposed by Dewe et al. (1998) and improved based on the feedback from a user study. The schema, shown in Fig. 4, consists of 7 main categories and 32 subcategories. The dataset was collected manually, which means that it is a designed

[3] The website as it was in 2015 can be viewed via Internet Archive Wayback Machine at https://web.archive.org/web/20150615095927/http://www.webgenrewiki.org/index.php5/Main_Page.

Main Category	Informational Description/ Explanation	Opinion	Narrative	Informational Persuasion	Interactive Discussion	How-To/ Instructional	Lyrical	Spoken
Sub-category	Course materials	Advertisement	Historical article	Description with intent to sell	Discussion forum	FAQ about how-to	Song lyrics	Formal speech
	Description of a person	Advice	Magazine article	Editorial	Reader/viewer responses	How-to	Poem	Interview
	Description of a thing	Letter to editor	News report/blog	Persuasive article or essay	Question/answer forum	Technical support	Prayer	TV/movie script
	Encyclopedia article	Opinion blog	Travel blog	Other	Other forum	Recipe	Other	Transcript of video/audio
	FAQ about information	Reviews	Personal blog			Other		Other
	Information blog	Religious blogs/sermons	Sports report					
	Legal terms and conditions	Other opinion	Short story					
	Technical report		Other narrative					
	Research article							
	Other information							

Fig. 5 The CORE schema as it is used in the CORE dataset (Egbert et al., 2015). The figure was created based on the information that was provided to us by the dataset authors

corpus, and consists of 1,280 texts that are prototypical for the assigned genre classes.

The *KRYS I* (Berninger et al., 2008) dataset consists of over 6,000 texts that are exclusively in the PDF format. The authors used their own hierarchical schema of 10 main categories and 70 subcategories, such as *Poetry Book, Menu, News Report, Contract, Essay, Sheet Music*. The dataset was collected manually by the students who collected instances of a genre that was assigned to them. Later, the instances were reclassified independently by two annotators.

The *SANTINIS-ML* dataset (Santini, 2010) is a multi-label extension of the 7-Web Genre Collection (Santini, 2007). Here, the authors introduced a double-layered genre schema with the aim of being able to cover all of the diversity of texts found on the web without having a very high granularity. The schema consists of four functional genres (*Descriptive-Narrative, Explicatory-Informational, Argumentative-Persuasive* and *Instructional*) and 11 specific genre categories – the labels from the 7-Web Genre Collection and 4 BBC web genres (*Editorials, DIY Mini-Guides, Short Biographies, Feature Articles*). The dataset consists of 2,480 texts, annotated by the author of the research.

The *I-EN-Sample* and *I-RU-Sample* (Sharoff, 2010) datasets consist of 250 texts, annotated with the Functional Genre Classification schema. The schema consists of 7 macro-genres: *Discussion, Information, Instruction, Propaganda, Recreation, Regulation*, and *Reporting*. They are defined solely based on the function or the purpose of the document and the 7 of them should be able to cover all the texts on the web. The novelty of this approach lies in the fact that the schema aims for a high coverage with a low granularity by using broad functional categories. The dataset was collected based on the results of random queries made to a search engine.

The *CORE* dataset (Egbert et al., 2015) is the result of one of the most extensive works on the genre schema and annotation process. The CORE schema was constructed following the bottom-up approach, where the end web users were asked to provide situational characteristics of web texts that would be relevant for them. The

Category Group	Objective Informative	Subjective Reporting	Opinion	Promotion	Dialogue	Literature	Formatted Text	Other
Category	News/Reporting	Opinionated News	Opinion/ Argumentation	Promotion	Interview	Script/Drama	FAQ	Other
	Announcement		Review	Promotion of a Product	Forum	Lyrical	List of Summaries /Excerpts	
	Information/ Explanation			Promotion of Services	Correspondence	Prose		
	Research Article			Invitation				
	Instruction							
	Recipe							
	Call							
	Legal/Regulation							

Fig. 6 The GINCO schema

initial schema was then improved based on a series of 10 pilot studies. The final schema as it is used in the dataset which is available online consists of 8 higher-level categories that cover the situational characteristics or functions of texts and 47 specific subcategories. The schema is shown in Fig. 5. A slightly modified schema with a smaller set of categories was used for annotation of Finnish (Laippala et al., 2019; Skantsi & Laippala, 2023), French and Swedish (Laippala et al., 2020; Repo et al., 2021) datasets, for which the detailed guidelines are available.[4] For the purposes of evaluation of a genre classifier model, 8 more datasets were recently annotated with CORE labels: an Indonesian dataset of around 1,000 texts, and 7 smaller evaluation datasets of 92 to 334 texts for Arabic, Catalan, Chinese, Hindi, Portuguese, Spanish and Urdu (see Laippala et al. (2022b)).

The CORE corpus consists of web texts that were extracted from the "General" part of the Corpus of Global Web-based English (GloWbE) (Davies & Fuchs, 2015). The GloWbE corpus was collected via Google searches with high frequency English 3-grams as the queries (Davies & Fuchs, 2015). The texts were then annotated via crowd-sourcing with 908 participants, where each text was annotated by 4 annotators. Despite leading the annotators through the decision process with a decision-tree survey, the inter-annotator agreement was shown to be low (see Sect. 4.2 for more details). The instances where there was no majority agreement on one label are annotated with multiple labels. The size of the dataset, as reported by the authors, is 53,000 texts, however, the dataset that is available online contains 48,420 texts. The dataset was used in further studies which provide a detailed analysis of linguistic characteristics of the CORE genres (Biber & Egbert, 2018), analyse the role of lexical and grammatical features in the performance and stability of genre classifiers (Laippala et al., 2021), or experiment with automatic genre classification (Ronnqvist et al., 2021; Repo et al., 2021).

The Slovene *Genre Identification Corpus (GINCO)* (Kuzman et al., 2022b) consists of 1,002 texts and uses a schema that is based on the subcategory level of the CORE schema, but also on other high-coverage schemata. In contrast to the CORE

[4] The guidelines for the annotation with the CORE schema are available here: https://turkunlp.org/register-annotation-docs.

	Principal Categories				Additional Categories
Category Group	**information**	**discussion**	**narration**	**promotion**	emotive
Categories	instruct	argum	fictive	compuff	flippant
	hardnews	scitech	personal	ideopuff	informal
	legal	eval		appell	specialist
	info				dialogue
					poetic

Fig. 7 The FTD schema

schema, it is not hierarchical and it has a lower granularity of categories to improve the inter-annotator agreement. The schema, shown in Fig. 6, consists of 23 specific genre categories and the category *Other*, dedicated to texts which do not fall into any other label. To alleviate the annotation, the genre categories were presented in category groups. The dataset was annotated by 2 expert annotators. They followed a multi-labelling approach where texts can be annotated with up to three genre labels. In this setting, the primary label is deemed to be the one that is most prevalent and the one that is used for single-label text classification, whereas the secondary and tertiary labels provide additional information on the fuzziness of the text, and the three levels can be used for multi-label classification. Extensive guidelines are available.[5] The dataset is a crawled corpus, as it is randomly sampled from two Slovenian web corpora from different time periods – the slWaC 2.0 corpus (Erjavec & Ljubešić, 2014) from 2014 and the MaCoCu-sl 1.0 corpus from a 2021 crawl (Bañón et al., 2022b). To be representative of the conditions on the web, it also includes noise, that is, a subset of "not suitable" texts, such as non-textual documents and automatically generated documents, multi-genre documents and other web-specific challenges.

The *Functional Text Dimensions* (FTD) dataset (Sharoff, 2018) in English and Russian follows another approach, introducing 18 Functional Text Dimensions (FTDs). As stated by the authors, the novelty of this approach is that "[u]nlike the traditional approach to designing genre lists or hierarchies in which texts need to be members of the respective classes, the FTD approach to detecting the genre of annotated texts is based on their similarity to prototypes, which is measured as distance in the FTD space" (Sharoff, 2018). The schema is a set of broad functional categories instead of specific labels, so that a text can be described through a combination of multiple labels, which could be regarded as parameters. By providing a possibility of assigning multiple labels to a text, the schema also deals with hybrid texts. In this approach, annotators define the presence of each dimension on a scale from 0 to 2, representing the distance of the text from the prototypical texts based on presence or absence of text features, typical for the genre class. As shown in Fig. 7, the schema consists of 12 principal categories, which means that a text needs to be given a non-zero score in at least one of them, and 6 additional categories, which describe additional variation between texts. While the principal categories

[5] The guidelines for annotation with the GINCO schema are available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

differentiate texts based on their function, additional categories provide information on other textual characteristics, for instance, on the stylistic differences, such as is the case of category *informal*, used for texts, written in informal language.

The FTD dataset consists of texts from multiple sources: the Pentaglossal corpus (5 g) (Forsyth & Sharoff, 2014), which consists of Russian and English texts with known origin (fiction, corporate communication, political debates, TED talks, UN reports, etc.), and a random selection of texts from the ukWac web corpus (Baroni et al., 2009) and the Russian GICR web corpus (Piperski et al., 2013). The FTD schema was extended to Arabic (Bulygin & Sharoff, 2018), a smaller set of FTDs was used for annotation experiments of additional English datasets (Suchomel, 2020), and a modified schema, consisting of 10 FTD labels, was used for annotation of a large Russian corpus, the *LiveJournal* dataset (Lepekhin & Sharoff, 2022). Guidelines for annotation are available.[6]

4.3.2 Information retrieval datasets

Information retrieval datasets were created with the aim to provide genre identification inside the search engine, so that the users would be able to find the texts more efficiently based on genre criteria. Thus, their schemata consist of categories that are deemed to be useful for the search engine users.

The *Genre-KI-04* (Eissen & Stein, 2004) dataset consists of 1,239 HTML files, annotated with a schema of 8 categories, devised based on a survey on genre usefulness (bottom-up approach). The schema consists of the following labels: *Help, Article, Discussion, Shop, Portrayal (non-priv), Portrayal (priv), Link Collection* and *Download*. Recently, the schema was used by Agrawal et al. (2019) for assessing the credibility of web pages.

The *7-Web Genre Collection* (Santini, 2007), in some works also referenced as SANTINIS, consists of 1,400 texts. It is annotated with 7 genre labels that are exclusive to the web: *Blog, e-Shop, FAQs, Online Newspaper Front Page, List, Personal Home Page*, and *Search Page*. The dataset, consisting of 200 instances of each genre, was manually collected by the author of the research, following the criteria of "annotation by objective sources".

The *20-Genre Collection* (MGC) (Vidulin et al., 2007) consists of 1,539 texts. They are classified into 20 genres: *Blog, Childrens', Commercial/Promotional, Community, Content Delivery, Entertainment, Error message, FAQ, Gateway, Index, Informative, Journalistic, Official, Personal, Poetry, Pornographic, Prose fiction, Scientific, Shopping*, and *User Input*. As the aim of the research was to construct a genre classifier which would identify genres within a search engine, the schema includes categories which would be of interest to the search engine users or which the users would want to filter out, e.g., the *Error message*. Based on the collection method, the corpus has both characteristics of a designed as well as a random corpus: first, instances were retrieved based on the searches made to the Google search

[6] Guidelines for annotation with the FTD schema are available here https://github.com/ssharoff/genre-keras/blob/master/annot-v4.md.

engine using popular keywords. Secondly, the collection was enlarged by adding random web pages, and finally, instances of less frequent genre categories were manually collected to obtain a balanced corpus. The collection was one of the firsts that opted for the multi-labelling approach. Recently, the schema was transformed into a hierarchical set of categories, which was shown to improve the automatic genre prediction (Madjarov et al., 2019).

4.3.3 Other datasets

The *Leeds Web Genre Corpus* (Asheghi et al., 2016) is one of the largest English genre datasets, consisting of almost 5,000 texts. The dataset consists of two subcorpora, created with the aim of evaluating the proposed method for manual annotation: a balanced, manually collected subcorpus, and a random subcorpus, collected based on the automated queries made to a search engine. The balanced subcorpus was used to test the reliability of the manual annotation on a more controlled set of texts, and then the annotation campaign was extended to the random subcorpus to test the method in realistic conditions. The authors devised their own schema, constructed following the top-down approach, starting from the Genre-KI-04 schema (Eissen & Stein, 2004). The schema consists of 20 specific genres, i.e., *Personal Homepage, Company/Business Homepage, Educational Organizational Homepage, Personal Blog/Diary, Online Shops, Instruction/How to, Recipe, News Article, Editorial, Conversational Forum, Biography, Frequently Asked Questions, Review, Interview, Story, Dictionary/Thesaurus Entries, Link Lists or Directories of Links, Song Lyrics, Quotes* and *Encyclopedic Articles*. The categories were shown to be able to cover 74% of the texts in the Leeds Corpus, while the remaining were annotated with the category *Other*. Despite the fact that the introduction of the category *Other* allows classification of any web text with the Leeds schema, the authors state that their main focus was not on "completeness of the genre inventory but on genre annotation methodology". Since the authors note that the aim behind the construction of the dataset and the corresponding schema is not to provide a high-coverage schema, but to analyse genre annotation, we do not include the Leeds Web Genre Corpus under high-coverage collections, but rather describe it separately. The dataset was annotated via crowd-sourcing where each text was annotated by 5 participants. The annotation campaign was shown to be successful and a high inter-annotator agreement was reported, with the percentage agreement of 88% and Fleiss' kappa (Fleiss, 1971) of 0.87 on the balanced subcorpus, and the percentage agreement of 78% and Fleiss' kappa of 0.71 on the random subcorpus.

5 Automatic genre identification

Genre researchers used various technologies for automatic genre identification, mostly depending on the state-of-the-art of machine learning algorithms at the time of research. In the past, support vector machines (SVMs) were most frequently used (Rezapour Asheghi, 2015; Laippala et al., 2017, 2021; Sharoff et al., 2010; Pritsos & Stamatatos, 2018; Petrenz & Webber, 2011), as it was shown that they are very

suitable for text categorization (Joachims, 1998). Other methods, previously used for genre classification, are the discriminant analysis (Feldman et al., 2009; Biber & Egbert, 2015) which is the earliest method that was applied to this task (Karlgren & Cutting, 1994), decision tree classifiers, more specifically the C4.5 algorithm (Finn & Kushmerick, 2006; Dewdney et al., 2001), Naive Bayes algorithm (Feldman et al., 2009; Priyatam et al., 2013) and graph-based models using hyper-link connections between web pages (Asheghi et al., 2014; Zhu et al., 2011). In recent studies, SVMs were used to obtain an insight into which feature sets are most relevant for genre identification (Laippala et al., 2021; Sharoff et al., 2010; Pritsos & Stamatatos, 2018; Asheghi et al., 2014). As noted by Laippala et al. (2021), it is very valuable to include linguistic analyses into the research to "not only achieve high performance but also to guarantee the validity of the results". For these further analyses, linear support vector machines were shown to be an appropriate choice.

Genre researchers experimented with various feature sets, e.g., bag-of-words, consisting of lexical features (words, word or character n-grams) and/or grammatical features (part-of-speech tags) (Sharoff, 2021; Laippala et al., 2021), punctuation, text statistics (Finn & Kushmerick, 2006), visual features of HTML web pages, such as HTML tags and images (Lim et al., 2005; Levering et al., 2008; Maeda & Hayashi, 2009), and URLs of web documents (Abramson & Aha, 2012; Jebari, 2014; Priyatam et al., 2013). However, results revealed that the discriminative features vary across studies and datasets, as well as across genre classes (see Laippala et al. (2021)). Nevertheless, multiple studies showed that lexical features describe genres better than grammatical ones, which was based on the comparison of lexical features and part-of-speech tags (Pritsos & Stamatatos, 2018; Laippala et al., 2021; Sharoff et al., 2010). Thus, when recent studies performed machine learning experiments with symbolic classifiers, that is, classifiers which are explainable and non-neural, they most commonly used lexical features – character and/or word n-grams (Pritsos & Stamatatos, 2018; Bulygin & Sharoff, 2018; Lepekhin & Sharoff, 2022) or a combination of lexical and grammatical features (Asheghi et al., 2014; Laippala et al., 2021). The latter was shown to provide better results than training on lexical information alone and to assure more stable models, that is, models which are capable of generalizing beyond the training data (see Laippala et al. (2021)). Recently, Kuzman and Ljubešić (2022) analysed the impact of using yet another type of grammatical features, namely syntactic dependencies, and showed that this textual representation is able to capture genre information better than lexical and part-of-speech features. In addition to this, when trained on syntactic dependencies, the classification model learns the structure of the sentences instead of word meanings, which quite likely assures that the prediction is not topic-dependent. This assumption, however, has not been proven yet.

Recent developments in NLP shifted the focus to neural networks due to significant improvements on all fronts. Researchers experimented with deep as well as shallow neural networks and found that for this task, the linear fastText (Joulin et al., 2017) model, which is a shallow neural model, is comparable to convolutional neural networks (CNN) with pre-trained word embeddings, while being computationally more efficient (see Laippala et al. (2019)). Furthermore, it was shown that the fastText model outperforms traditional methods, i.e., SVMs,

decision trees, random forest classifiers, logistic regression classifiers and the Naive Bayes classifier at this task (Kuzman & Ljubešić, 2022).

Recently developed Transformer-based pre-trained language models led to a significant breakthrough in this field. Repo et al. (2021) and Kuzman et al. (2022b) showed that these deep learning models can achieve strong performance on small amount of training data, and, what is even more, they can reach around 30 points higher micro and macro F1 scores than the previous best model, fastText (see Kuzman et al. (2022b)). One likely reason for such drastic improvements obtained through the Transformer models in comparison to previous methods could be the fact that Transformer text representations incorporate information on syntax as well, which was shown to be very informative for genre classification (Kuzman and Ljubešić (2022)). Moreover, the models are capable of good levels of cross-lingual transfer. This was demonstrated when zero-shot cross-lingual experiments were performed by learning on a large English CORE dataset and testing on smaller datasets in Finnish, French, and Swedish (Repo et al., 2021). The cross-lingual experiments resulted in F1 scores between 0.61 and 0.69, while the results of training and testing the classifier on the same datasets in a monolingual setting range between 0.73 and 0.83 F1 (see Table 4). This was further researched by Ronnqvist et al. (2021) who showed that by combining all the four CORE datasets into a multilingual dataset, the performance is further improved, reaching micro F1 scores between 0.72 and 0.84. The multilingual classifier also achieved better zero-shot performance, as the classifier trained on three of the datasets and tested on the fourth reached micro F1 scores between 0.63 and 0.80. The highest scores were achieved with the multilingual XLM-RoBERTa model (Conneau et al., 2020) which outperformed other multilingual pre-trained Transformer language models in multiple genre studies (Ronnqvist et al., 2021; Repo et al., 2021; Kuzman & Pollak, 2022a). Furthermore, the XLM-RoBERTa model was revealed to be comparable (Kuzman et al., 2022b) or better (Repo et al., 2021) than monolingual models. This is in contrast to other NLP tasks, where the monolingual models were shown to outperform the multilingual models (Ulčar et al., 2021).

In addition to choosing the optimal features and machine learning algorithms to achieve good classification performance, genre researchers need to tackle additional genre-specific challenges, such as handling noise, i.e., texts which do not fit under any of the categories, and dealing with hybrid texts by performing multi-label or span-based classification. Handling noise was researched by Pritsos and Stamatatos (2018), and regarding the hybrid texts, Santini (2006) proposed a new method of multi-label classification: classifying texts by applying several classification models on them, trained on different genre datasets with different genre schemata. Recently, multi-label classification was also taken up by Madjarov et al. (2019), who performed multi-label classification and hierarchical multi-label classification using Predictive Clustering Trees (PCTs), Sharoff (2021), who used a Bi-directional Long Short-Term Memory (BiLSTM) classifier (Yogatama et al., 2017), by Ronnqvist et al. (2021), who used multilingual Transformer models for cross-lingual multi-label classification experiments, and Laippala et al. (2022a) who used a monolingual Transformer model for multi-label classification experiments.

Table 3 Accuracy obtained on the best-performing feature (for each of the datasets), reported in the experiments with the SVM (Sharoff et al., 2010)

Dataset	Accuracy
7-Web Genre Collection (Santini, 2007)	97.14
Genre-KI-04 (Eissen & Stein, 2004)	85.81
Hierarchical genre collection (Stubbe & Ringlstetter, 2007)	65.72
KRYS I (Berninger et al., 2008)	62.02
I-EN-Sample (Sharoff, 2010)	60.40
20-Genre Collection (Vidulin et al., 2007)	56.45

Another challenge for genre identification is assuring stability of the trained models. If the genre-annotated dataset is not general enough, the models could learn to classify texts based on other characteristics that are common to the texts of one class, such as the topic, instead of genre characteristics, and would not be able to generalize beyond the dataset (Sharoff et al., 2010). Petrenz and Webber (2011) assessed the stability of various machine learning methods by analysing their performance across changes in topic-genre distribution and advised that "stability should join accuracy as a criterion for assessing any new developments in genre classification". This was further explored by Laippala et al. (2021) who analysed which features provide the most stable SVM models. Since the advent of deep neural models, exploration of their stability has been tightly connected with research on their explainability, as the models act as black boxes. Thus, to explore which genres are more affected by topical biases, Lepekhin and Sharoff (2021) performed adversarial attacks on Transformer models, and Ronnqvist et al. (2022) introduced an Integrated Gradients input attribution method to achieve stable explanations of genre predictions based on keyword lists. The issue of topical shift which impacts the performance of the models was recently addressed by Lepekhin and Sharoff (2022) who performed experiments on two genre-annotated datasets with different genre distribution. The results showed that training the Transformer models on multiple datasets improves the performance. In addition to that, they experimented with ensembles of a monolingual Russian RuBERT (Kuratov & Arkhipov, 2019) model, multilingual XLM-RoBERTa (Conneau et al., 2020) and the Logistic Regression classifier and showed that ensembles of pre-trained models mostly outperform single Transformer models.

As genre-annotated datasets were created with different end goals and as no reference genre-annotated dataset exists, previous machine learning experiments focusing on automatic genre identification are "self-contained, and corpus-dependent" (Rehm et al., 2008). In addition to using different datasets, genre classes, classification algorithms and feature sets, researchers reported performance of the models with different metrics, making comparison between the experiments impossible. The last comparison of genre-annotated datasets based on AGI experiments was published by Sharoff et al. (2010) who compared most of the relevant genre datasets that existed at the time of research: the Hierarchical genre collection (Stubbe & Ringlstetter, 2007), the I-EN-Sample (Sharoff, 2010), Genre-KI-04 (Eissen & Stein,

Table 4 Overview of the reported results from the most relevant in-dataset experiments on recently developed genre datasets

Research	Dataset	Number of labels	Model	Best score (per dataset)
Asheghi et al. (2014)	Leeds corpus (Asheghi et al., 2016)	15 (subset of Leeds labels)	Semi-supervised multi-class min-cut graph-based algorithm (Ganchev & Pereira, 2007)	Accuracy: 0.90
Sharoff (2021)	FTD (English), FTD (Russian) (Sharoff, 2018)	10 (subset of FTD labels)	BiLSTM (Yogatama et al., 2017)	Precision: 0.77, 0.75
Repo et al. (2021)	FinCORE, FreCORE, SweCORE, CORE (Egbert et al., 2015)	7 (subset of CORE main labels)	XLM-RoBERTa large (Transformer model) (Conneau et al., 2020)	F1: 0.73, 0.77, 0.83, 0.76
Kuzman et al. (2022b)	GINCO	12 (merged GINCO labels)	SloBERTa (Transformer model) (Ulčar & Robnik-Šikonja, 2021)	Micro F1: 0.70, Macro F1: 0.67
Laippala et al. (2022a)	CORE (Egbert et al., 2015)	56 (CORE main and sub-categories; multilabel approach)	BERT large (Transformer model) (Kenton & Toutanova, 2019)	Micro F1: 0.68

2004), the KRYS I (Berninger et al., 2008) dataset, the 20-Genre Collection (Vidulin et al., 2007) and the 7-Web Genre Collection (Santini, 2007) dataset. They used a linear SVM model, and the part-of-speech n-grams, character and word n-grams as the features. The results that were obtained on the best-performing feature set, shown in Table 3, show very high results obtained by the classifiers trained on the 7-Web Genre Collection or the Genre-KI-04 dataset. While they performed with accuracy over 85, the results for classifiers trained on other datasets ranged between 56 to 66 in terms of accuracy. The worst results were obtained on the 20-Genre Collection. Despite the high results of some of the classifiers, the authors noted that "these impressive results might not be transferable to the wider web" due to the fact that the collections are not comparable between each other even on similar categories, that they are lacking in representativeness and that the annotations either cannot be tested for reliability or were revealed to not be sufficiently reliable. This was confirmed via cross-dataset experiments, that is, by testing the classifiers on all other compared datasets. There, the scores for accuracy were mostly below 40.

The results from the machine learning experiments on the genre datasets that were developed since then are reported in Table 4. The reported results give the reader a feel of the current state-of-the-art performance of genre classifiers on the datasets, however, as the experiments are corpus- and technology-dependent, the results are hardly comparable. Furthermore, as can be seen from Table 4, most of the experiments used only a subset of labels from their high-coverage schemata, and thus the results do not show the trained classifiers' capacity of identifying genres in the entire dataset. Thus, to be able to compare the suitability of genre datasets for automatic genre identification, a study is needed which would include all of the datasets.

First step towards comparing recent genre datasets was made by Kuzman and Pollak (2022a) who explored the comparability of the Slovene GINCO dataset (Kuzman et al., 2022b) and the English CORE dataset (Egbert et al., 2015). They mapped the original GINCO categories and CORE subcategories to a joint schema and performed cross-dataset experiments. As the datasets are in different languages, the experiments were cross-lingual at the same time. The experiments showed that the datasets are comparable enough to allow cross-dataset and cross-lingual transfer, reaching micro and macro F1 scores between 0.60–0.64 and 0.52–0.66. For comparison, the in-dataset results ranged between 0.78–0.81 in micro F1 and 0.73–0.84 in macro F1 for GINCO, and reached 0.77 micro F1 and 0.72 macro F1 for CORE. The results are comparable to the cross-lingual experiments, performed by Repo et al. (2021) on the English, Swedish, French and Finnish datasets, annotated with the CORE schema. Both studies also revealed that despite the fact that the CORE dataset has up to 40-times more instances than other compared datasets, the difference in size does not result in significantly better results. This might indicate that the pre-training of Transformer models gives such informative representations that Transformer models reach high results already after a few thousand instances observed, and adding tens of thousands of instances to the training does not significantly improve the results. This hypothesis is supported by recent experiments on the CORE dataset, which showed that the learning curve of a Transformer language model started flattening already after using 30% of training data (around 10,000

texts) (Laippala et al., 2022a). In addition, the study revealed that the beginning of the document is the most informative part for genre prediction, which means that the Transformer models are potent enough to only need a short part of a text to recognize a genre.

Another theory could be that the CORE dataset is less suitable for the automatic genre identification, since the reported low inter-annotation agreement (Egbert et al., 2015; Sharoff, 2018) could mean that there is more noise in the data. However, as performance of genre classifiers was often improved by training on multiple datasets (Lepekhin & Sharoff, 2022; Ronnqvist et al., 2021), the next step could be to experiment with training a classifier on the combination of GINCO and CORE, since they were shown to be comparable. The described research explored comparability to some extent, but it should be noted that since the labels vary significantly in terms of granularity, the joint schema did not cover all of the CORE and GINCO labels. In addition to this, the experiments were single-label. Thus, texts belonging to discarded labels, and CORE texts annotated with more than one class were not used in the experiments. This means that the comparison was performed on reduced datasets.

Classification experiments based on a joint schema seem hardly possible without any interventions to the data due to many differences between most relevant high-coverage datasets: the CORE (Egbert et al., 2015), FTD (Sharoff, 2018) and GINCO (Kuzman et al., 2022b) datasets. The FTD dataset is annotated using the dimensional and not categorical approach, meaning that the texts are annotated with multiple labels. Similarly, the CORE corpus is annotated with two levels of categories, main categories and subcategories, where some texts belong to multiple main categories, multiple subcategories or to no subcategory at all. This further complicates mapping to a joint schema, and requires multi-label experiments. However, an analysis of the real usefulness of the datasets and their schemata for classification applied to new data would be very beneficial to the community. To this end, an indirect comparison via downstream evaluation could be performed, either on samples of various web corpora, or on whole web corpora. Each of the datasets could be used to train a separate classifier, using the original schema. The classifiers would be applied to the sample, or the whole corpus. The sample could be used for estimating the accuracy and usefulness of the predictions directly by human annotators, while the whole annotated corpus could be used in various extrinsic evaluations, and evaluated on the target task.

Despite the challenges in automatic genre identification that still exist, the recent technology breakthrough has allowed creation of Transformer-based genre classifiers which can already be applied to large new datasets in numerous languages. Recently, Laippala et al. (2022b) published the Register Oscar dataset, consisting of 351 million documents in 14 languages to which genre labels were automatically assigned. The dataset was created by applying the multilingual genre classifier (Ronnqvist et al., 2021) on the OSCAR datasets (Suárez et al., 2019). The classifier is trained on the CORE datasets and uses the 8 main CORE labels. The zero-shot classification was evaluated based on annotated datasets in eight new languages (Arabic, Catalan, Chinese, Hindi, Indonesian, Portuguese, Spanish and Urdu). The results for the new languages were promising, ranging between 0.58 and 0.82 F1 scores.

Similarly, the genre classifier based on the GINCO dataset (Kuzman et al., 2022b) is planned to be applied to massive monolingual and parallel web corpora, produced for 10 under-resourced European languages in the scope of the MaCoCu project (Bañón et al., 2022a). The MaCoCu corpora, automatically annotated with genre labels, are reported to be released in 2023.[7] As opposed to genre datasets, presented in Sect. 4, which were manually annotated with genre classes and used as the basis for training AGI classifiers, the Register Oscar and MaCoCu datasets are automatically annotated with genre classes and are ones of the first results of applying genre classifiers on unseen collections of texts.

6 Conclusion

This survey presents an exhaustive overview of previous research on automatic genre identification. Thereby, it uncovers numerous challenges of the task and the approaches to dealing with them, further explained in the following paragraphs:

1. Lack of consensus on the main concepts, connected with the automatic genre identification, leading to a proliferation of competing genre schemata and approaches.
2. Existence of hybrid texts and multi-text documents in web datasets, related to the limited control over the content creation and collection of web texts, which pose challenges for manual annotation and automatic classification.
3. Lack of comparability between existing genre datasets, aggravated by the fact that they are hard to find or have ceased to be available.
4. Challenges, connected with assuring stability of genre classifiers and their good performance also in out-of-domain scenarios, so that the classifiers are not dataset-dependent.

The first challenge is that there exists no consensus on the term that should be used for this phenomenon, on the definition of the concept itself, and on the set of labels to be included into the schema to assure high coverage, reliable annotation and good classification results. As researchers define genres differently and perform research on them with different aims, they use very different genre schemata. We analysed 16 of them and showed that out of 177 labels only 48 labels are used in more than one schema.

Secondly, the reliability of genre datasets and consequently also the results of automatic genre identification can be negatively impacted by the fundamental difficulties connected with the genre notion itself. During annotation and machine learning experiments, researchers need to deal with hybrid texts, i.e., texts which have features of multiple genres, texts without any discernible purpose or features, and documents that consist of multiple texts, such as a letter, posted as a part of a blog. On the web, writers are not constrained to following genre norms, thus, the texts

[7] https://macocu.eu/

vary significantly in regard of their genre prototypicality. These difficulties contributed to rather low inter-annotator agreement in some of the annotation campaigns, resulting in genre datasets with questionable reliability.

Thirdly, the last review of existing genre datasets, published in 2010 (see Sharoff et al. (2010)), exposed the lack of comparability, representativeness and reliability of existing genre datasets, and highlighted a need for further research on AGI. They called for efforts to develop a large reference corpus, collected from a diverse range of sources, and with a genre schema which would allow for reliable annotation. At that time, genre datasets almost exclusively contained only English documents, thus, the authors also urged for a creation of datasets in other languages, which would enable cross-lingual automatic genre identification. Our paper is the first work that surveys the research on automatic genre identification that was done in the last 12 years, describing that research in the context of the older work as well.

In this period, the technological progress delivered crawling tools which enable fast collection of thousands, or even millions of texts from the web, assuring representativeness of the datasets. Secondly, the researchers benefited from the advent of crowd-sourcing platforms, which allow quick and relatively inexpensive annotation of tens of thousands of texts. And finally, recently developed deep neural machine learning technologies relying on self-supervision, such as the Transformer language models, were shown to be able to classify genres drastically better than any previous technology, even when trained on only a thousand of texts. Furthermore, the models were revealed to achieve good results on cross-lingual genre classification which opens the doors to empowering a multitude of languages with the benefits of automatic genre identification.

While early genre datasets mostly focused on a small set of specific labels, recently, numerous datasets were developed with schemata which aim to cover the entire genre composition of the web, which is the first step to providing a general dataset with labels that could be applied to any corpus. Such schemata are the FTD schema (Sharoff, 2018), based on which English, Russian and Arabic genre datasets were created, and the CORE schema (Egbert et al., 2015) which served as a basis for English, Finnish, French, Swedish, Slovene and other datasets. These datasets addressed the lack of representativeness, comparability and reliability that was observed for earlier genre collections.

While recently many new genre-annotated datasets emerged, previous datasets ceased to be available. This paper is the first to provide an extensive list of all available genre datasets with information on where to access them. Despite a recent positive trend of publishing genre datasets, the majority of datasets were hard to find. They are not published in a certified repository which would assure reliable long-term archiving. This should be addressed by the community to assure that the datasets are available even 10 years after the research was published. In addition to this, to facilitate further advances in this field, researchers should strive to also publish the code used in their AGI experiments, to make the research reproducible and open.

Recent emergence of multiple high-coverage datasets in various languages has been an important step towards developing a genre classifier which could be applied to any dataset in multiple languages. Recent research (Ronnqvist et al., 2021; Lepekhin & Sharoff, 2022) achieved very promising results by training Transformer

models on multiple datasets joined together. Thus, the key to a general cross-lingual genre identifier could lay in merging comparable high-coverage datasets. By training a model on multiple datasets, it would be less prone to topical and other biases, which would assure its stability. To be able to merge the high-coverage datasets, i.e., the CORE dataset (Egbert et al., 2015), the FTD dataset (Sharoff, 2018) and the GINCO dataset (Kuzman et al., 2022b), a cross-dataset analysis is necessary. To avoid interfering with the datasets and schemata, for each of the datasets, a separate Transformer model could be trained, using its original schema. Then each classifier could be applied to all other high-coverage datasets. This would result in each text from each of the datasets being labelled three times: with a GINCO category, a CORE category and an FTD category. The quantitative analysis of the results would reveal comparability of the genre schemata. Moreover, it could be extended to an extrinsic evaluation to analyse the usability of assigned genre labels for the users. Based on this, two promising directions are possible: 1) merging the datasets into one based on a joint schema to obtain a more stable and potent genre classifier; or 2) instead of creating just one model and applying it on a dataset, texts could be automatically annotated with predictions of all three classifiers, the GINCO-based, FTD-based and CORE-based classifier, allowing the corpora users to choose the labels that are most relevant for their use case.

Development of general and stable genre classifiers is a first step to providing reliable automatic genre identification. However, as observed by previous work, the field needs more research that would address additional challenges related to this task, i.e., performing multi-label classification on hybrid texts, handling noise, and separating multiple texts in a multi-text document. Only then we will have classifiers that will be truly useful in realistic conditions on the web.

Author contributions All authors contributed to the study conception and design. Literature search, data collection and analysis were performed by TK. The first draft of the manuscript was written by TK and NL, who also critically revised the work. All authors read and approved the final manuscript.

Funding This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019-2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

Availability of data and materials The authors confirm that all datasets analysed during this study are referenced in this published article and the links to their location are provided.

Code availability Not applicable.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Abramson, M., & Aha, D.W. (2012). What's in a URL? Genre Classification from URLs. Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Agrawal, S., Sanagavarapu, L.M., & Reddy, Y.R. (2019). FACT-Fine grained assessment of web page CredibiliTy. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 1088–1097.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. In: First International Workshop on Innovative Information Systems, pp. 85–92.

Asheghi, N.R., Markert, K., & Sharoff, S. (2014). Semi-supervised graph-based genre classification for web pages. In: Proceedings of TextGraphs-9: The Workshop on Graph-Based Methods for Natural Language Processing, pp. 39–47.

Asheghi, N. R., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation, 50*(3), 603–641.

Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., & Suchomel, V. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 301–302.

Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., & Zaragoza, J. (2022). Slovene web corpus MaCoCu-sl 1.0. (Slovenian language resource repository CLARIN.SI)

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation, 43*(3), 209–226.

Berninger, V.F., Kim, Y., & Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. BCS-IRSG Workshop on Corpus Profiling, pp. 1–10.

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press.

Biber, D., & Egbert, J. (2015). Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science, 2*(1), 3–36.

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Boese, E.S. (2005). Stereotyping the web: Genre classification of web documents (Unpublished doctoral dissertation). Citeseer.

Bulygin, M., & Sharoff, S. (2018). Using machine translation for automatic genre classification in Arabic. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 153–162.

Chandler, D. (1997). An introduction to genre theory.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V.,Wenzek, G., Guzmán, F., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451.

Crowston, K., Kwaśnik, B., & Rubleske, J. (2010). *Problems in the use-centered development of a taxonomy of web genres. Genres on the Web* (pp. 69–84). Springer.
Davies, M. (2004). British National Corpus (from Oxford University Press). Available online at https://www.english-corpora.org/ bnc/
Davies, M. (2008). The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/ coca/
Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1–28.
Dewdney, N., Van Ess-Dykema, C., & MacMillan, R. (2001). The form is the substance: Classification of genres in text. In: Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management.
Dewe, J., Karlgren, J., & Bretan, I. (1998). Assembling a balanced corpus from the internet. In: Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998), pp. 100–108.
Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology, 66*(9), 1817–1831.
Erjavec, T., & Ljubešić, N. (2014). The slwac 2.0 corpus of the slovene web. T. Erjavec, J. Žganec Gros (ur.). *Jezikovne tehnologije zbornik, 17*, 50–55.
Feldman, S., Marin, M.A., Ostendorf, M., & Gupta, M.R. (2009). Part-of-speech histograms for genre classification of text. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4781–4784.
Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology, 57*(11), 1506–1518.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.
Forsyth, R. S., & Sharoff, S. (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing, 29*(1), 6–22.
Freund, L., Clarke, C.L., & Toms, E.G. (2006). Towards genre classification for IR in the workplace. In: Proceedings of the 1st International Conference on Information Interaction in Context, pp. 30–36.
Ganchev, K., & Pereira, F. (2007). Transductive structured classification through constrained min-cuts. In: Proceedings of the Second Workshop on Textgraphs: Graph-Based Algorithms for Natural Language Processing, pp. 37–44.
Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: Proceedings of the Fifth Web as Corpus Workshop, pp. 27–35.
Jebari, C. (2014). A pure URL-based genre classification of web pages. In: 2014 25th International Workshop on Database and Expert Systems Applications, pp. 233–237.
Jebari, C. (2021). Enhancing the identification of web genres by combining internal and external structures. *Pattern Recognition Letters, 146*, 83–89.
Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning* (pp. 137–142). Springer.
Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the Fifteen Conference of the European Chapter of the Association for Computational Linguistics, 2*, 427–431.
Kanaris, I., & Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. *IEEE International Conference on Tools with Artificial Intelligence, 2*, 3–10.
Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management, 45*(5), 499–512.
Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of the 15th International Conference on Computational Linguistics.
Kennedy, A., & Shepherd, M. (2005). Automatic identification of home pages on the web. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 99c–99c.
Kenton, J.D.M.-W.C., & Toutanova, L.K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacl-hlt, pp. 4171–4186.
Kilgarriff, A. (2012). Getting to know your corpus. In: International Conference on Text, Speech and Dialogue, pp. 3–15.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.
Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.
Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 333–339.
Kuzman, T., & Ljubešić, N. (2022). Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In D. Mladenić & M. Grobelnik (Eds.), *Odkrivanje znanja in podatkovna skladišča - SiKDD: 10.* Institut Jožef Stefan.
Kuzman, T., Rupnik, P., & Ljubešić, N. (2022). The GINCO training dataset for web genre identification of documents out in the wild. *Proceedings of the language resources and evaluation conference* (pp. 1584–1594). European Language Resources Association.
Kuzman, T V. N., & Pollak, S. (2022). Assessing comparability of genre datasets via cross-lingual and cross-dataset experiments. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: Zbornik konference* (pp. 100–107). Institute of Contemporary History.
Kwaśnik, B. H., & Crowston, K. (2005). *Introduction to the special issue: Genres of digital documents*. Information Technology & People.
Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics, pp. 292–297.
Laippala, V., Luotolahti, J., Kyröläinen, A.-J., Salakoski, T., & Ginter, F. (2017). Creating register subcorpora for the Finnish Internet Parsebank. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 152–161.
Laippala, V., Rönnqvist, S., Hellström, S., Luotolahti, J., Repo, L., Salmela, A., & Pyysalo, S. (2020). From web crawl to clean register-annotated corpora. In: Proceedings of the 12th Web as Corpus Workshop, pp. 14–22.
Laippala, V., Salmela, A., Rönnqvist, S., Aji, A.F., Chang, L.-H., Dhifallah, A., & Skantsi, V. (2022). Towards better structured and less noisy web data: Oscar with register annotations. In: Proceedings of the eighth workshop on noisy user-generated text (w-nut 2022), pp. 215–221.
Laippala, V., Egbert, J., Biber, D., & Kyröläinen, A.-J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation, 5*, 1–32.
Laippala, V., Rönnqvist, S., Oinonen, M., Kyröläinen, A.-J., Salmela, A., Biber, D., & Pyysalo, S. (2022). Register identification from the unrestricted open web using the corpus of online registers of English. *Language Resources and Evaluation, 1*, 1–35.
Lee, Y.-B., & Myaeng, S.H. (2002). Text genre classification with genrerevealing and subject-revealing features. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 145–150.
Lee, Y.-B., & Myaeng, S.H. (2004). Automatic identification of text genres and their roles in subject-based categorization. In: 37th Annual Hawaii International Conference on System Sciences.
Lee, D. (2002). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Teaching and learning by doing corpus analysis* (pp. 245–292). Brill Rodopi.
Lepekhin, M., & Sharoff, S. (2021). Experiments with adversarial attacks on text genres. arXiv preprint arXiv:2107.02246
Lepekhin, M., & Sharoff, S. (2022). Estimating confidence of predictions of individual classifiers and their ensembles for the genre classification task. *Proceedings of the language resources and evaluation conference* (pp. 5974–5982). European Language Resources Association.
Levering, R., Cutler, M., & Yu, L. (2008). Using visual features for fine-grained genre classification of web pages. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), pp. 131–131.
Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management, 41*(5), 1263–1276.
Lukin, A., Moore, A.R., Herke, M., Wegener, R., & Wu, C. (2011). Halliday's model of register revisited and explored.
Madjarov, G., Vidulin, V., Dimitrovski, I., & Kocev, D. (2019). Web genre classification with methods for structured output prediction. *Information Sciences, 503*, 551–573.

Maeda, A., & Hayashi, Y. (2009). Automatic genre classification of Web documents using discriminant analysis for feature selection. In: 2009 Second International Conference on the Applications of Digital Information and Web Technologies, pp. 405–410.

Mason, J.E., Shepherd, M., & Duffy, J. (2009). An n-gram based approach to automatically identifying web page genre. In: 2009 42nd Hawaii International Conference on System Sciences, pp. 1–10.

Moessner, L. (2001). Genre, text type, style, register: A terminological maze? *European Journal of English Studies, 5*(2), 131–138.

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021). Genre as Weak Supervision for Cross-lingual Dependency Parsing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4786–4802.

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly, 5*, 541–574.

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics, 37*(2), 385–393.

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., & Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In: Proceedings of 8th Web as Corpus Workshop (WAC-8), pp. 24–29.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora (Unpublished doctoral dissertation)*. Masaryk university Faculty of informatics.

Pritsos, D., & Stamatatos, E. (2018). Open set evaluation of web genre identification. *Language Resources and Evaluation, 52*(4), 949–968.

Priyatam, P. N., Iyengar, S., Perumal, K., & Varma, V. (2013). Don't use a lot when little will do: Genre identification using URLs. *Research in Computing Science, 70*, 233–243.

Rehm, G. (2002). Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, pp. 1143–1152.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., & Vidulin, V. (2008). *Towards a reference corpus of web genres for the evaluation of genre identification systems*. Lrec.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., & Laippala, V. (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In: 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, EACL 2021, pp. 183–191.

Rezapour Asheghi, N. (2015). *Human annotation and automatic detection of web genres (Unpublished doctoral dissertation)*. University of Leeds.

Rönnqvist, S., Kyröläinen, A.-J., Myntti, A., Ginter, F., & Laippala, V. (2022). Explaining Classes through Stable Word Attributions. Findings of the association for computational linguistics: Acl 2022, pp. 1063–1074.

Rönnqvist, S., Skantsi, V., Oinonen, M., & Laippala, V. (2021). Multilingual and zero-shot is closing in on monolingual web register classification. In: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 157–165.

Rosso, M. A. (2008). User-based identification of Web genres. *Journal of the American Society for Information Science and Technology, 59*(7), 1053–1072.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., & Liu, X. (2001). Genre based navigation on the web. In: Proceedings of the 34th annual Hawaii international conference on system sciences, p. 10.

Santini, S.M. (2006). Common criteria for genre classification: Annotation and granularity. In: Workshop on Text-based Information Retrieval (TIR-06). Conjunction with ECAI 2006, Riva del Garda, 2006.

Santini, M. (2007). *Automatic identification of genre in web pages (Unpublished doctoral dissertation)*. University of Brighton.

Santini, M. (2010). *Cross-testing a genre classification model for the web. Genres on the Web* (pp. 87–128). Springer.

Santini, M., Mehler, A., & Sharoff, S. (2010). *Riding the rough waves of genre on the web. Genres on the Web* (pp. 3–30). Springer.

Sharoff, S. (2021). Genre annotation for the web: text-external and textinternal perspectives. Register studies.

Sharoff, S. (2010). *In the garden and in the jungle genres on the web* (pp. 149–166). Springer.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora, 13*(1), 65–95.

Sharoff, S., Wu, Z., & Markert, K. (2010). *The Web Library of Babel: Evaluating genre collections*. Lrec.
Shavrina, T. (2019). Genre classification problem: In pursuit of systematics on a big webcorpus. *Proceedings of Third Workshop Computing, 4*, 70–83.
Skantsi, V., & Laippala, V. (2023). Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics, 1*, 1–31.
Snow, R., O'connor, B., Jurafsky, D., & Ng, A.Y. (2008). Cheap and fast– but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254–263.
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471–495.
Stein, B., Eissen, S. M. Z., & Lipka, N. (2010). *Web genre analysis: Use cases, retrieval models, and implementation issues Genres on the Web* (pp. 167–189). Springer.
Stewart, J. G., & Callan, J. (2009). *Genre oriented summarization (Unpublished doctoral dissertation)*. Language Technologies Institute, School of Computer ScienceCarnegie Mellon University.
Ströbel, M., Kerz, E., Wiechmann, D., & Qiao, Y. (2018). Text genre classification based on linguistic complexity contours using a recurrent neural network. MRC@ IJCAI, pp. 56–63.
Stubbe, A., & Ringlstetter, C. (2007). Recognizing genres. Towards a reference corpus of web genres: Proceedings.
Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell Oxford.
Suárez, P.J.O., Sagot, B., & Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In: 7th Workshop on the Challenges in the Management of Large Corpora (cmlc-7).
Suchomel, V. (2020). Genre Annotation of Web Corpora: Scheme and Issues. In: Proceedings of the Future Technologies Conference, pp. 738–754.
Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.
Ulčar, M., Žagar, A., Armendariz, C.S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. arXiv preprint arXiv:2107.10614 .
Van der Wees, M., Bisazza, A., & Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
Vidulin, V., Luštrek, M., & Gams, M. (2007). Using genres to improve search engines. In: 1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing, pp. 45–51.
Williams, M., & Crowston, Kevin. (2000). Reproduced and emergent genres of communication on the World WideWeb. *Information Society, 16*(3), 201–215.
Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. In: Thirty-fourth International Conference on Machine Learning (ICML 2017).
Zhu, J., Zhou, X., & Fung, G. (2011). Enhance web pages genre identification using neighboring pages. In: International Conference on Web Information Systems Engineering, pp. 282–289.
Zu Eissen, S.M., & Stein, B. (2004). Genre classification of web pages. In: Annual Conference on Artificial Intelligence, pp. 256–269.

2.5 Final Remarks

In this chapter, we identified gaps in prior work and outlined best practices for developing high-coverage genre datasets and classifiers, based on the findings of the journal article *Automatic Genre Identification: A Survey* (Kuzman & Ljubešić, 2023) that is enclosed in Section 2.4. We identified several challenges related to the development of a genre schema, manually-annotated genre datasets, and genre classifiers, which are addressed in this thesis.

First, our review of existing genre schemata and datasets showed that none can be directly applied to our study. The existing high-coverage schemata provide a good starting point, however, they have several limitations: too high label granularity that negatively impacts the reliability of manual annotations, limited comparability across schemata that prevents possible merging of multiple existing genre datasets; abstract label names that may not be understood by end users; and closed label sets that do not account for instances falling outside predefined categories. To address these issues, we set to develop a new genre schema that overcomes the limitations of current high-coverage schemata. This schema would be based on the existing schemata to allow for potential mapping between genre schemata and integration of existing genre datasets into our machine learning experiments. This is particularly important given the lack of comparability between datasets in previous work, which has led to dataset-specific experiments and results that offer limited insight into the robustness of the developed genre classifier, that is, its generalization capabilities when applied to other datasets.

Second, no existing high-coverage dataset completely fulfills the requirements of our study. We identified two freely-accessible genre dataset families that use high-coverage schemata, namely the CORE (Egbert et al., 2015; Laippala et al., 2019; Laippala et al., 2020; Repo et al., 2021; Skantsi & Laippala, 2023) and FTD datasets (Sharoff, 2018). However, neither includes Slovenian language which is one of the focuses of our study. Creating a new manually-annotated dataset based on the CORE and FTD schemata would be challenging, as previous attempts to apply these schemata to new datasets have reported low inter-annotator agreement. To ensure more reliable annotation, we aim to propose a new genre schema and an improved annotation methodology. To ensure that the Slovenian genre dataset represents real-world web texts well, we will follow best practices from previous work, including the extraction of random instances from web corpora and attention to temporal diversity.

Third, although recent studies report strong performance of genre classifiers, they did not publish the models. As a result, we have to develop our own genre classifiers. Fine-tuned BERT-like Transformer-based models were shown to be the most promising machine learning method to achieve a robust performance, that is, to achieve generalization across datasets and languages. They will serve as our starting point to develop multilingual genre classifiers that can be applied on Slovenian, English and other European languages. To assess the generalization capabilities of the model, it should be evaluated in the cross-dataset scenario where it is applied to different datasets. In order for this evaluation to be feasible, a mapping across genre schemata needs to be developed that does not only take into account the similarity of label names – it must be based on a deep understanding of the involved datasets and their instances, since the same label names may be defined in annotation guidelines differently and be therefore represented by very different texts.

Chapter 3

Development of Robust Genre Schemata and Datasets

The examination of prior research, presented in Chapter 2, indicates that although existing genre schemata, datasets, and best practices offer a useful basis for our research on automatic genre identification, they are not directly applicable to our study. Consequently, this chapter outlines the development of novel genre schemata, manually-annotated datasets, and genre classifiers as part of our goal to develop a robust multilingual genre classifier that demonstrates high coverage when applied on web texts.

In this chapter, we address the following goals:

- Firstly, we fulfill Goal *G1* "Develop a high-coverage genre schema that encompasses the full spectrum of genre variation in web texts and ensures reliable manual annotation with high inter-annotator agreement." in Sections 3.1 and 3.2.1, in which we present the GINCO and X-GENRE genre schemata.

- Secondly, we fulfill Goal *G2* "Propose a novel methodology for manual genre annotation that ensures high reliability based on a novel genre schema, detailed annotation guidelines, and continuous training of expert annotators." in Section 3.1.2, in which we present the annotation campaign and annotation guidelines for the GINCO schema. This allows us to test Hypothesis *H1*: "*Acceptable inter-annotator agreement in genre identification annotation campaigns can be accomplished through a manual annotation approach that is based on a well-designed genre schema, comprehensive annotation guidelines, and employment of expert annotators.*", which is supported by the results.

- Lastly, we fulfill the first part of Goal *G3* "Develop a genre dataset for Slovenian, which will be the first of its kind for this language, to enable training and evaluation of machine learning models in automatic genre identification on Slovenian language. Additionally, we will construct training datasets that include English, [...]", as we present in Sections 3.1.3 and 3.2.2 the Slovenian manually-annotated GINCO dataset and the Slovenian-English manually-annotated X-GENRE dataset, respectively.

This chapter is organized as follows. In Section 3.1 we introduce the GINCO genre schema (Section 3.1.1), propose a refined annotation methodology and develop comprehensive annotation guidelines (Section 3.1.2). Additionally, we create a new Slovenian training dataset for automatic genre identification, named the GINCO dataset (Section 3.1.3). We conduct machine learning experiments to assess the suitability of the GINCO schema and dataset for the development of genre classifiers, and conclude with suggestions for further improvement of the proposed schema (Section 3.1.4). Second, in Section

3.2 we explore the integration of the Slovenian GINCO dataset with pre-existing English manually-annotated genre datasets. We propose a unified schema to facilitate the merging of the schemata from the existing English datasets with the GINCO schema (Section 3.2.1), resulting in the creation of a new Slovenian-English dataset (Section 3.2.2). Finally, we compare the two newly proposed datasets with high-coverage genre datasets from previous studies through in-dataset experiments, demonstrating that the newly developed multilingual genre dataset provides superior training data for the development of genre classifiers (Section 3.2.2).

3.1 The Slovenian GINCO Dataset with a Novel Genre Schema and Annotation Guidelines

In Chapter 2, a key gap in the existing literature is identified: the absence of a manually-annotated genre dataset that includes Slovenian, the language of interest in this study. Therefore, such a dataset must be created from scratch. Given the limitations of existing genre schemata and annotation approaches, discussed in Sections 2.1 and 2.2, it is necessary to develop an improved genre schema and annotation guidelines. This will ensure that the new manually-annotated dataset contains reliable annotations which enable its use for the development and evaluation of genre classifiers.

This section introduces a novel genre schema, referred to as the GINCO schema (Section 3.1.1), an annotation methodology designed to ensure high inter-annotator agreement (Section 3.1.2), and a new Slovenian training dataset for automatic genre identification, named the GINCO dataset (Section 3.1.3). Finally, in Section 3.1.4, multiple machine learning models are applied to the GINCO dataset to evaluate its suitability as a training and testing resource for genre classifiers, as well as to provide preliminary insights into the models' potential performance on this task in the Slovenian language.

This section is based on the accompanying paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al., 2022). For further details on the schema, dataset, methodology, experiments, and results, refer to the paper that is enclosed in Section 3.1.5.

3.1.1 GINCO Genre Schema

In this section, we propose a new genre schema, addressing Goal *G1* of this thesis, which aims to develop a high-coverage genre schema that encompasses the full spectrum of genre variation in web texts and ensures reliable manual annotation with high inter-annotator agreement. A suitable schema is a crucial foundation for ensuring reliable manual annotation and satisfactory performance of genre classifiers. When designing the schema, we established the following criteria to be met:

- High-coverage: The label set should provide sufficient coverage to enable classification of the majority of texts in a web collection under specific labels.
- Low granularity: It is important to maintain a limited number of labels, as increased granularity can negatively affect both the reliability of manual annotations and the performance of models.
- Intuitive label names: Our objective is for the final genre-annotated datasets to be accessible to a wider audience, enabling users to filter data by genre and use the resulting genre-specific datasets for machine learning applications or linguistic research. Consequently, the proposed genre labels should be easily interpretable by end users.

The proposed schema, henceforth referred to as the GINCO schema, builds on previous genre schemata to ensure comparability with existing high-coverage schemata, and to leverage insights from prior research. Specifically, to enable potential cross-lingual experiments using the large manually-annotated English genre dataset CORE (Egbert et al., 2015), our proposed schema is based on the subcategories of the CORE schema (Egbert et al., 2015). The original CORE schema comprises approximately 50 subcategories, which are in some cases highly similar and not delineated well, as noted by Biber and Egbert (2018). Thus, as a first step, we merge similar CORE subcategories to decrease schema granularity. Secondly, to increase schema coverage and improve comparability with other schemata, we incorporate labels from additional schemata (Asheghi et al., 2016; Boese, 2005; Crowston et al., 2010; Laippala et al., 2020; Rehm et al., 2008; Roussinov et al., 2001; Stubbe & Ringlstetter, 2007). When selecting additional labels, we avoid abstract names in order to ensure that the final genre labels are easily understood by the general public.

Category Group	Objective Informative	Subjective Reporting	Opinion	Promotion	Dialogue	Literature	Formatted Text	Other
Category	News/Reporting	Opinionated News	Opinion/ Argumentation	Promotion	Interview	Script/Drama	FAQ	Other
	Announcement		Review	Promotion of a Product	Forum	Lyrical	List of Summaries /Excerpts	
	Information/ Explanation			Promotion of Services	Correspondence	Prose		
	Research Article			Invitation				
	Instruction							
	Recipe							
	Call							
	Legal/Regulation							

Figure 3.1: The GINCO genre schema.

The final GINCO schema, shown in Figure 3.1, consists of 24 labels – 23 specific genre categories and an additional category *Other*. Following the best practices suggested by Asheghi et al. (2016), the inclusion of the *Other* category addresses the limitation of previous high-coverage genre schemata, which assume that all web texts can be classified within a closed set of genre labels. This category is intended for texts that either lack distinct genre characteristics or do not fit under any of the specific genre labels.

The GINCO genre categories are defined based on the purpose, conventional form, and lexico-grammatical features typical of each genre. Detailed descriptions of the categories are provided in the appendix of the paper by Kuzman, Rupnik, et al. (2022) in Section 3.1.5, and in the annotation guidelines (see Section 3.1.2).

3.1.2 Improved Annotation Methodology

A manual annotation campaign can be considered successful if the final annotations are found to be reliable. Reliability is assessed through inter-annotator agreement, which indicates whether the annotators share a common understanding of the labels. Previous work has shown that achieving reliable manual annotation for the task of automatic genre identification is very challenging (Egbert et al., 2015; Sharoff, 2018; Suchomel, 2020; Zu Eissen & Stein, 2004). In response, we propose an improved annotation methodology to ensure high annotation reliability, which corresponds to Goal *G2* of this thesis.

The first step toward making the annotation task more manageable for annotators is developing a genre schema with a reasonable granularity and clearly defined categories, as

discussed in Section 3.1.1. In this section, we describe additional approaches implemented to achieve acceptable inter-annotator agreement for this task, described in detail in the accompanying paper by Kuzman, Rupnik, et al. (2022), enclosed in Section 3.1.5. The main improvements to the annotation guidelines include the adoption of objective annotation criteria, the creation of detailed guidelines supported by a decision tree to guide the annotators through the decision process, and the involvement of expert annotators who receive continuous training.

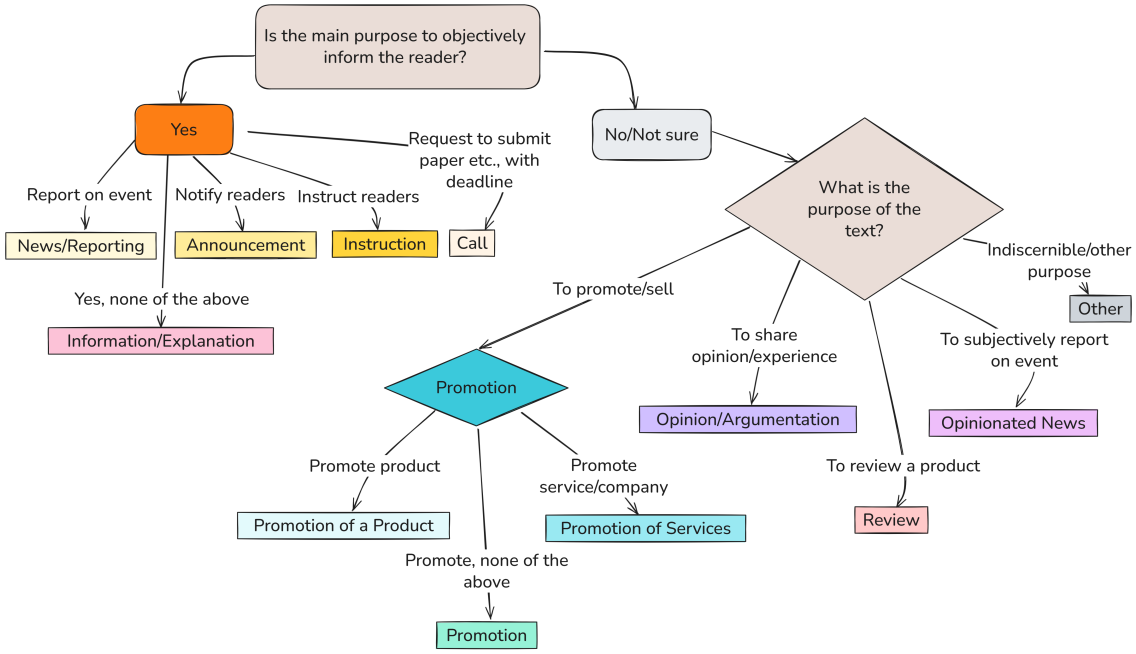


Figure 3.2: A subset of the GINCO annotation decision tree.

Firstly, to ensure a common understanding of genre labels, we dedicate considerable effort to the development of comprehensive annotation guidelines. The guidelines consist of two main components: (1) a decision tree, and (2) detailed descriptions of each genre category. The decision tree, shown in Figure 3.2, assists annotators in navigating the decision-making process through a series of questions and criteria. The descriptions of genre categories, depicted in Figure 3.3, describe the purpose, text form, and typical lexico-grammatical features, associated with each genre class. To further support annotators, the guidelines provide multilingual examples and recommendations for addressing challenging cases. The annotation guidelines are publicly accessible to facilitate future manual annotation campaigns.[1]

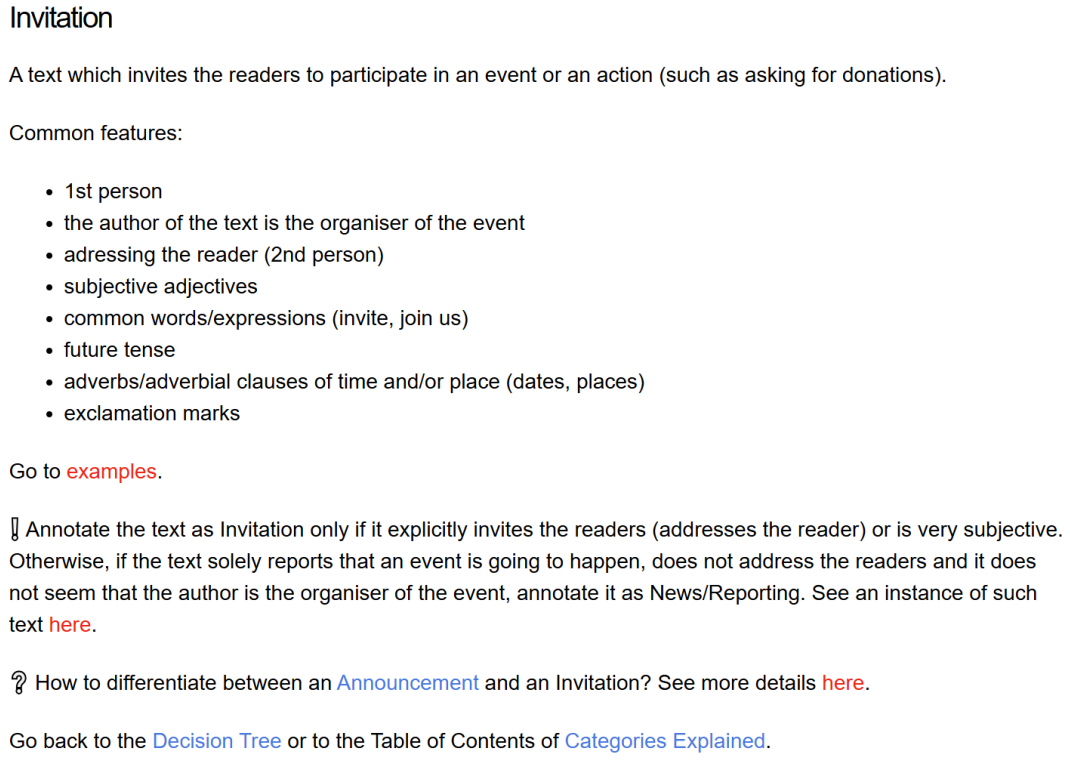
Particular attention is given to ensuring that the criteria in both the decision tree and label descriptions are as objective as possible. Unlike the FTD (Sharoff, 2018) genre annotation approach, which relies on "the impression the annotators obtained from reading a text" (Sharoff, 2018), we instruct the annotators to base their decisions on the presence of concrete lexico-grammatical features in the text. This approach prevents the annotators from primarily relying on intuition or subjective judgment, which would compromise annotation reliability and decrease inter-annotator agreement.

Finally, previous annotation campaigns, such as the CORE annotation approach (Egbert et al., 2015), relied on crowd-sourcing, which presents specific limitations, including

[1]The GINCO annotation guidelines can be accessed at https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

restricted control over annotators and a lack of information regarding their reliability. In contrast, we employ expert annotators, specifically two PhD candidates in computational linguistics with a background in linguistics. They received initial training and continuous guidance. The annotators regularly discussed hard cases and instances on which they initially disagreed in order to reach a shared understanding of the labels and to ensure final annotations that are consistent and of high quality.

To assess the usefulness of the proposed GINCO schema and the improved annotation methodology, we calculate inter-annotator agreement on the labels of the two annotators assigned to the manually-annotated GINCO dataset prior to their discussions. Following Krippendorff (2018), a nominal Krippendorff's alpha of 0.667 is considered as the threshold for acceptable agreement. In our study, the inter-annotator agreement reached a Krippendorff's alpha of 0.71, which is above the acceptable threshold. This indicates that the proposed genre schema and annotation methodology enable reliable manual annotation for genre identification. Furthermore, our approach offers improved annotation reliability compared to the CORE schema (Egbert et al., 2015), for which a Krippendorff's alpha of 0.53 across 56 subcategories and 0.66 for 8 main categories was reported (Sharoff, 2018). These results are consistent with Hypothesis *H1*: *Acceptable inter-annotator agreement in genre identification annotation campaigns can be accomplished through a manual annotation approach that is based on a well-designed genre schema, comprehensive annotation guidelines, and employment of expert annotators.*



Invitation

A text which invites the readers to participate in an event or an action (such as asking for donations).

Common features:

- 1st person
- the author of the text is the organiser of the event
- adressing the reader (2nd person)
- subjective adjectives
- common words/expressions (invite, join us)
- future tense
- adverbs/adverbial clauses of time and/or place (dates, places)
- exclamation marks

Go to examples.

Annotate the text as Invitation only if it explicitly invites the readers (addresses the reader) or is very subjective. Otherwise, if the text solely reports that an event is going to happen, does not address the readers and it does not seem that the author is the organiser of the event, annotate it as News/Reporting. See an instance of such text here.

How to differentiate between an Announcement and an Invitation? See more details here.

Go back to the Decision Tree or to the Table of Contents of Categories Explained.

Figure 3.3: An example of a genre category description from the GINCO annotation guidelines.

3.1.3 Slovenian Manually-Annotated Dataset GINCO

The annotation campaign, presented in Section 3.1.2, is applied to a Slovenian text collection to develop the first Slovenian genre dataset, known as the Genre Identification Corpus GINCO 1.0 (Kuzman et al., 2021). With this dataset, we accomplish the first part of Goal *G3* that focuses on the development of a genre dataset for the Slovenian language. To adhere to the principles of findability, accessibility, interoperability, and reusability (FAIR) (Wilkinson et al., 2016) and ensure long-term availability, the GINCO dataset is made publicly accessible via CLARIN.SI and Hugging Face data repositories.[2]

The development of the GINCO dataset involves two phases: the collection of representative Slovenian text instances, and the manual annotation proposed in Section 3.1.2. To construct the dataset, random samples of text instances are extracted from crawled web corpora. This approach has been shown to produce datasets that more accurately reflect real-world web content compared to manually collected designed corpora (Asheghi et al., 2016). Particular care is taken to ensure that the dataset includes instances from a wide range of sources covering diverse topics, in order to prevent potential biases in genre classifiers trained on the dataset, which could be related to specific topics, authors, or web domains. Accordingly, the GINCO dataset comprises samples from two Slovenian web corpora collected during different time periods: 501 texts from the slWaC 2.0 corpus (Erjavec & Ljubešić, 2014) from 2014, and 501 texts from the MaCoCu-sl 1.0 corpus from 2021 (Bañón, Esplà-Gomis, Forcada, Garcıéa-Romero, et al., 2022). The final GINCO dataset consists of 1,002 Slovenian web texts, amounting to nearly 500,000 words. These texts are manually-annotated with 24 genre categories according to the GINCO schema. The dataset is divided into 602 training instances, 200 evaluation instances, and 200 test instances, with the distribution of labels maintained consistently across these splits. Further details on the GINCO dataset are provided in the paper by Kuzman, Rupnik, et al. (2022), enclosed in Section 3.1.5.

3.1.4 Machine Learning Experiments on the GINCO Dataset

To evaluate the suitability of the GINCO dataset for automatic genre identification, we use it as a training and test dataset in a series of machine learning experiments, involving a shallow neural fastText model (Joulin et al., 2017), and two Transformer-based BERT-like models: the monolingual Slovenian SloBERTa model (Ulčar & Robnik-Šikonja, 2021a, 2021b) and the multilingual XLM-RoBERTa model (Conneau et al., 2020). The set of experiments is described in detail in the paper by Kuzman, Rupnik, et al. (2022) in Section 3.1.5.

Table 3.1: Performance of classifiers trained and evaluated on the GINCO dataset. Results are reported in terms of micro-F1 and macro-F1 scores. The best scores for each metric are marked in bold.

Model	Micro-F1	Macro-F1
stratified dummy	0.067	0.061
fastText	0.361 ± 0.007	0.219 ± 0.013
XLM-RoBERTa	0.624 ± 0.015	**0.579 ± 0.024**
SloBERTa	**0.629 ± 0.016**	0.575 ± 0.037

As shown in Table 3.1, the fine-tuned Transformer-based language models achieve

[2]The GINCO dataset is accessible on the CLARIN.SI repository at http://hdl.handle.net/11356/1467 and on the Hugging Face repository at https://huggingface.co/datasets/cjvt/ginco.

micro-F1 scores between 0.62 and 0.63, and a macro-F1 of 0.58. We regard these results as promising given the small amount of training data. However, it should be noted that some previous studies reported higher scores. Laippala et al. (2022) achieved a micro-F1 of 0.68 using CORE subcategories on the CORE dataset (Egbert et al., 2015), while Repo et al. (2021) reported F1 scores between 0.73 and 0.83 using fine-tuned BERT-like models on the FinCORE (Laippala et al., 2019), FreCORE, SweCORE (Laippala et al., 2020), and CORE (Egbert et al., 2015) datasets, using the seven main CORE labels (see Section 2.3).

The analysis of label-level performance, detailed in the paper by Kuzman, Rupnik, et al. (2022) in Section 3.1.5, suggests that further refinements of the schema might be needed to boost model performance. To explore this, we experiment with reducing the number of labels by merging related categories, resulting in a simplified schema with 12 genre labels.

Table 3.2: The effect of reducing the GINCO schema granularity on the SloBERTa model performance. Statistical testing in each column is performed with the Mann-Whitney U rank test with p-value encoded with asterisks: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

Schema	Micro-F1	Macro-F1
Initial GINCO schema	0.629 ± 0.016	0.575 ± 0.037
Simplified GINCO schema (12 labels)	$\mathbf{0.696 \pm 0.011}$ ***	$\mathbf{0.668 \pm 0.028}$ ***

As shown in Table 3.2, the reduction in the schema granularity leads to a significant improvement in the performance of the SloBERTa model, achieving a micro-F1 of 0.70 and a macro-F1 of 0.67. Following these promising results, further improvements to the schema, particularly in terms of granularity, are explored in the following sections (see Section 3.2).

3.1.5 Related Paper: The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1584–1594
Marseille, 20-25 June 2022

The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild

Taja Kuzman, Peter Rupnik, Nikola Ljubešić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
taja.kuzman@ijs.si, peter.rupnik@ijs.si, nikola.ljubesic@ijs.si

Abstract

This paper presents a new training dataset for automatic genre identification GINCO, which is based on 1,125 crawled Slovenian web documents that consist of 650,000 words. Each document was manually annotated for genre with a new annotation schema that builds upon existing schemata, having primarily clarity of labels and inter-annotator agreement in mind. The dataset consists of various challenges related to web-based data, such as machine translated content, encoding errors, multiple contents presented in one document etc., enabling evaluation of classifiers in realistic conditions. The initial machine learning experiments on the dataset show that (1) pre-Transformer models are drastically less able to model the phenomena, with macro F1 metrics ranging around 0.22, while Transformer-based models achieve scores of around 0.58, and (2) multilingual Transformer models work as well on the task as the monolingual models that were previously proven to be superior to multilingual models on standard NLP tasks.

Keywords: automatic genre identification, web genres, genre classification schema, web corpora, Slovenian language

1. Introduction

With the arrival of the Web, it has become significantly easier to collect very large corpora that fuel innovation and creation of advanced resources and language technologies. However, contrary to the traditionally collected corpora, web corpora are built in an automated way which limits the control over the contents that constitute the final corpus (Baroni et al., 2009). One of the post hoc evaluation methods to investigate the corpus composition and quality, and to enrich the corpus with important metadata is automatic genre identification (AGI). This method focuses on genres as text categories based on the author's purpose, the socially recognized function of a document and/or the conventional patterns of form, following the definition by Orlikowski and Yates (1994).

As this research is a part of the MaCoCu[1] project that aims to collect large corpora for under-resourced languages, our main purpose is to provide a classifier that would efficiently identify genres in web corpora for Slovenian and other languages, which would allow an in-depth analysis of the quality and composition of the newly provided corpora.

In addition to this, annotating the data with genre is beneficial in many other areas where language technology can be improved by a more fine-grained document typology, such as in part-of-speech tagging (Giesbrecht and Evert, 2009), zero-shot dependency parsing (Müller-Eberstein et al., 2021), automatic summarization (Stewart and Callan, 2009), and machine translation (Van der Wees et al., 2018). Furthermore, numerous genre annotation studies have been conducted with the aim of improving the Information Retrieval (IR) tools (Stubbe and Ringlstetter, 2007; Zu Eissen and Stein, 2004; Vidulin et al., 2007; Roussinov et al., 2001; Finn and Kushmerick, 2006; Boese, 2005).

Our contributions in this work are as follows. First, we propose a new genre classification schema, based on the previous schemata but modified with the goal of 1. using labels recognizable to the corpora users and 2. achieving high inter-annotator agreement, based on considering lexico-grammatical characteristics in addition to the purpose and form of the text. Second, we present the Genre Identification Corpus – GINCO 1.0 (Kuzman et al., 2021), a realistic dataset randomly sampled from two Slovenian web corpora that also includes noise, multi-genre documents and other web-specific challenges, enabling evaluation of classifiers in realistic conditions. Finally, we perform machine learning experiments over the new dataset and share some interesting insights with the community.

2. Related work

Despite the considerable benefits of automatic genre identification (AGI), no established classification exists (Sharoff, 2010). The genre researchers are not consistent in the use of terminology, and they refer to genres, text types, functional text dimensions or registers in different ways (Sharoff, 2018; Egbert et al., 2015; Laippala et al., 2021; Lee, 2002). Furthermore, there is no consensus on the genre definition. Consequently, most studies use their own genre schema, either hierarchical (Stubbe and Ringlstetter, 2007; Egbert et al., 2015) or not (Asheghi et al., 2016; Sharoff, 2018), and applying single (Santini, 2007; Sharoff, 2010) or multiple label annotation (Vidulin et al., 2007; Laippala et al., 2019). Schemata vary significantly regarding the number of classes as well, which range from seven (Santini, 2007; Sharoff, 2010; Lee and Myaeng, 2002) to more than hundred (Roussinov et al., 2001) and almost 300

[1] https://macocu.eu/

classes (Crowston et al., 2010). Consequently, the annotated corpora are not comparable, and testing of the similarity between them using cross-classification results in low accuracy (see Sharoff et al. (2010)).

Various approaches used to tackle genre identification revealed the task to be challenging due to fundamental difficulties emanating from the genre notion itself. Firstly, the conventions that characterise a genre category are not fixed or static, and instances of genre vary in their prototypicality (see Santini et al. (2010), Santini (2006)). Secondly, web documents sometimes display features of more than one genre, not fitting in discrete classes (see Sharoff (2021) and Repo et al. (2021)). Example of such hybrid texts is a promotion of a product written in a form of a news article. Thirdly, the purpose of the communication cannot always be discerned (see Williams (2000)).

The schemata that serve as a basis of the most recent genre identification studies are the schema of the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Functional Text Dimensions (FTD) approach (Sharoff, 2018) for English and Russian. The approaches were extended to cross-lingual genre classification experiments to eliminate the need for time-consuming and expensive manual annotation of large corpora in other languages. Bulygin and Sharoff (2018) performed genre classification on an Arabic web corpus using machine-translated English and Russian corpora annotated with FTDs. Using smaller manually annotated Finnish, Swedish, and French corpora, Repo et al. (2021) demonstrated that good levels of cross-lingual transfer from the extensive English CORE corpus to other languages can be achieved through a zero-shot learning setting performed with the Transformer-based pre-trained language models. Furthermore, the research showed that these models can achieve strong performance monolingually on small training data. A subsequent study (Rönnqvist et al., 2021) further improved the zero-shot results by performing zero-shot classification with multilingual pre-trained language models, trained on genre corpora in all four languages.

3. Dataset construction

3.1. Corpora

We performed genre annotation on two Slovenian web corpora, crawled in different time periods, the slWaC 2.0 corpus (Erjavec and Ljubešić, 2014) from 2014 and the MaCoCu-sl 1.0 corpus from 2021 (Bañón et al., 2022). The corpora were collected by crawling the Slovenian top-level domain (TLD) `.si`, as well as some generic-domain websites highly interlinked with the national TLD. The two corpora have the same structure, and were compiled and preprocessed using the same machinery, i.e., the SpiderLing[2] crawler (Suchomel and Pomikálek, 2012), the jusText[3] tool for boilerplate removal (Pomikálek, 2011), and the onion[4] tool for identifying duplicates (Pomikálek, 2011). In these web corpora, we consider two paragraphs to be near-duplicates if their intersection of word 5-grams exceeds the 50% threshold. To circumvent spurious topic-genre correlations, and to represent the distribution of genres on the Slovenian internet as closely as possible, we used a random selection of texts.

Documents not deemed to be suitable for genre annotation were labelled with the following *Not Suitable* categories: *Machine Translation*, *Generated Text*, *Not Slovene*, *Encoding Issues*, *HTML Source Code*, *Boilerplate*, *Too Short/Incoherent*, *Too Long* (longer than 5,000 words), *Non-Textual* (no full sentences, e.g., tables, lists), and *Multiple Texts* (multiple texts that cannot be split). This preprocessing step resulted in the "not suitable" part of the Genre Identification Corpus GINCO, which contains 123 texts. Interestingly, although the two Slovenian corpora were equally represented in the dataset, 89% of the unsuitable texts were from the recently crawled corpus and are mostly *Non-textual*, *Machine Translation* or *Too Short/Incoherent*, which points to the conclusion that the quality of Slovenian web texts has deteriorated since the web crawl in 2014. Additionally, the documents that consisted of multiple texts of different genres, where one text is followed by another, so they can be separated, such as news article, followed by comments, were split accordingly into two or more texts. They were assigned IDs that provide information on their origin and the order of the texts, so that they can be further analysed or merged back. At the end, the final dataset – the "suitable part" of GINCO on which the annotation was performed – consisted of 501 texts from the 2014 corpus and 501 texts from the 2021 corpus, i.e., 1002 texts in total.

While we did not perform any experiments in automating the splitting of documents containing multiple texts, our initial experiments in discriminating between suitable and unsuitable texts showed for the problem to be rather hard, achieving a macro F1 of 0.715, while the random baseline macro F1 is around 0.5. At this point, given the class imbalance (123 unsuitable texts vs. 1002 suitable texts), more unsuitable texts are classified as suitable than unsuitable, which we consider not to be satisfactory. Both tasks – eliminating noise from the web corpora and splitting documents containing texts of different genres – are kept for future work.

3.2. Annotation schema

The construction of the annotation schema was based on the following goals:

1. To reach high coverage with respect to real world corpora – to this end, we avoid using schemata that focus on a small set of specific genres (Zu Eissen and Stein, 2004; Santini, 2006; Asheghi et

[2] `http://corpus.tools/wiki/SpiderLing`
[3] `http://corpus.tools/wiki/Justext`
[4] `http://corpus.tools/wiki/Onion`

al., 2016; Lee and Myaeng, 2002; Boese, 2005). Instead, we propose category groups, based on main communicative purposes, identified in previous research (Egbert et al., 2015; Sharoff, 2010; Santini, 2010; Sharoff, 2018).

2. To consider usability for the corpora users and to provide genre labels, recognizable to users as much as possible – to this aim, we avoid abstract labels, such as *content delivery* (Vidulin et al., 2007), *resources* (Stubbe and Ringlstetter, 2007), *recreation* (Sharoff, 2010), *commpuff* (Sharoff, 2018).

3. To provide categories that the annotators can understand with little training and on which they largely agree – to this end, we avoid schemata with a very fine granularity (Crowston et al., 2010; Roussinov et al., 2001; Williams, 2000; Egbert et al., 2015; Berninger et al., 2008) which usually increases the ambiguity (see Sharoff (2010)).

Considering the prospect of extending the research to cross-lingual transfer from large English genre annotated corpus CORE (Egbert et al., 2015) to smaller languages, proposed by Repo et al. (2021) and Rönnqvist et al. (2021), we base our annotation schema on the CORE schema which is hierarchical and consists of 8 higher-level categories and more than 50 subcategories. We focus on the subcategories, for which "the labels are intuitive and correspond to register categories in other corpora" (Laippala, 2019). However, the inter-annotator agreement (see Egbert et al. (2015)) and linguistic description of the categories (see Biber and Egbert (2018)) revealed that there is room for improvement. Firstly, due to high granularity of the schema, the results revealed low inter-annotator agreement, as there was no majority agreement, i.e., agreement between at least three of four annotators, on the main category of 31.06% of documents and on the subcategory of 48.98% documents. For at least 10 subgenres there were no instances with majority agreement. Additionally, Biber and Egbert (2018) found that some subcategories are not well-defined linguistically, and that some are highly similar. Thus, we reduce the granularity by merging some of similar categories, such as *News Report* and *Sport Report*, and by including some less frequent categories into broader ones. Secondly, in an attempt to encompass as many relevant web genres as possible, we include some additional genres, identified in a survey of 200 random documents from the slWaC 2.0 Slovenian web corpus (Erjavec and Ljubešić, 2014), which were considered in previous schemata (Roussinov et al., 2001; Boese, 2005; Laippala et al., 2020; Stubbe and Ringlstetter, 2007; Asheghi et al., 2016; Crowston et al., 2010; Rehm et al., 2008).

Our schema does not group categories in the 8 main categories proposed by CORE (*Narrative*, *Opinion*, *Informational Description/Explanation*, *Interactive Discussion*, *How-to/Instructional*, *Informational Persuasion*, *Lyrical*, *Spoken*), since Biber and Egbert (2018) noted that some main categories encompass subgenres that are situationally, linguistically, and functionally different, that some similar subgenres are spread across different main categories and that some main categories overlap extensively and could be merged. Thus, 23 genre categories, based on recognizable labels, purpose, conventional forms and lexico-grammatical features, are grouped into 7 category groups (see Figure 1), 6 of which are based on purpose (*Objective Informative*, *Subjective Reporting*, *Opinion*, *Promotion*, *Dialogue*, *Literature*), whereas 1 is based solely on the form of the text (*Formatted Text*), consisting of visually very easily recognizable genres that could have various purposes, such as *Frequently Asked Questions*. The category groups were introduced to alleviate the annotation decision process, whereas the annotation and machine learning experiments were performed on the level of categories only. For more details on the categories, see their descriptions in the Appendix 1.

Unlike most schemata which consider the case where all web pages should belong to a predefined taxonomy of genres (Lim et al., 2005; Santini, 2007; Sharoff, 2018), we recognize that there might exist web pages that would not fall into any of the predefined genre labels. Following Asheghi et al. (2016), to deal with such web pages, we introduce the 24th category *Other*, intended for texts which purpose is unknown or not covered by other labels. We suggest that during the annotation process, the annotators keep record of possible additional genres that are annotated as *Other*, and if their presence in the corpora is significant, they can be added to the category set.

3.3. Annotation procedure

The genre annotation was conducted individually by two annotators – the author of the annotation schema and a second annotator. Both annotators are PhD students in the field of computational linguistics and have a linguistic background. The second annotator received short training on examples that were deemed to be prototypical, and was provided with a decision-tree survey which the annotators used until they sufficiently familiarized themselves with the task.

In addition to this, the annotators followed detailed annotation guidelines with examples of prototypical texts and descriptions of the purpose, form and common lexico-grammatical linguistic characteristics of the genres[5]. With the latter, we diverge from the approaches of the FTD (Sharoff, 2018) and CORE (Egbert et al., 2015) studies. They avoid considering the linguistic characteristics of genres in the annotation process and base the annotation on "the impression the annotators obtained from reading a text" (Sharoff,

[5]The guidelines for multiple languages are available at `https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/`.

Category Group	Objective Informative	Subjective Reporting	Opinion	Promotion	Dialogue	Literature	Formatted Text	Other
Category	News/Reporting	Opinionated News	Opinion/ Argumentation	Promotion	Interview	Script/Drama	FAQ	Other
	Announcement		Review	Promotion of a Product	Forum	Lyrical	List of Summaries /Excerpts	
	Information/ Explanation			Promotion of Services	Correspondence	Prose		
	Research Article			Invitation				
	Instruction							
	Recipe							
	Call							
	Legal/Regulation							

Figure 1: The Genre Schema

2018) to allow further analyses of linguistic characteristics of genre categories without circularity. We argue that we cannot disregard the possibility that based on annotators' prior knowledge and experience with genres, their decision may be influenced by linguistic characteristics nevertheless. Additionally, by instructing the annotators to base their decisions on presence of concrete characteristics in the text rather than on their impression of the text, the decisions are less subjective, which improves the inter-annotator agreement and consistency. The key linguistic characteristics were chosen based on a preliminary study of 200 texts from the Slovenian web corpora, and although they were not based on the linguistic analyses of the English CORE corpora by Biber and Egbert (2018), they happened to largely correspond to them.

In contrast to the annotation process of CORE (Egbert et al., 2015) which used single-labelling approach, we opted for multi-labelling approach where texts can be annotated with up to three genre labels. In this setting, the primary label is deemed to be the one that is most prevalent and the one that is mainly used for the automated identification, whereas the secondary and tertiary labels provide additional information on the fuzziness of the text, known as a hybrid.

3.3.1. Inter-annotator agreement

The annotation was performed in 21 batches of, on average, 50 texts, and the two annotators discussed uncertain cases in meetings, with frequency of meetings decreasing with progression of the annotation campaign. The inter-annotator agreement was calculated on the annotators' labels that were assigned prior to the discussions. The nominal Krippendorff's alpha (Krippendorff, 2018), calculated for agreement on the complete set of 24 categories on the level of primary labels only, reached 0.71, which is above the acceptable threshold of 0.67, defined by Krippendorff (2018). This shows that the proposed schema and annotation procedure improve the reliability of annotation in comparison with the CORE schema, for which Sharoff (2018) reported nominal Krippendorff's alpha of 0.53 on the complete set of 56 subcategories and 0.66 on the 8 main categories. Regarding the FTD approach, initial research (Sharoff, 2018) achieved higher agreement, reaching Krippendorff's alpha above 0.76, but subsequent research (Suchomel, 2020) reported significantly lower results, i.e., nominal alpha of 0.497, despite using only 9 of initial 18 FTD categories.

Additionally, we performed a manual analysis of agreement where we considered two cases as partial agreement: agreement between the primary label assigned by one annotator with the secondary label assigned by the other, and the agreement between the two primary labels that are a part of the same category group. The analysis revealed perfect agreement on primary labels in 73%, partial agreement in 19%, and no agreement in 8% of texts. Further analysis of cases with no agreement revealed that in 47% of such cases, no agreement was observed due to difficult categorisation of texts with features of many different genres or none at all. Secondly, in 32% of cases there was no agreement due to instances of new phenomena that were not previously sufficiently covered by the annotation guidelines. This shows that as the authors of web texts demonstrate varying levels of expertise or willingness to conform to the conventions, defined by genre experts, (see Sharoff (2021)) annotation guidelines and the schema based on previous work and small preliminary analyses cannot entirely cover all of the diversity found on the Web. Thus, updating the guidelines (and schema) as the task progresses is recommended.

3.4. Dataset encoding and availability

The final dataset, Genre Identification Corpus GINCO 1.0 (Kuzman et al., 2021) [6], consists of the "suitable" and "not suitable" subset, which are released separately in form of a JSON file. The suitable subset consists of texts, manually labelled with genres, and the non-suitable subset comprises documents, considered to be noise and labelled with *Not Suitable* categories. The corpus contains additional metadata, i.e., URL, domain, year, and attribute `hard`, indicating whether a

[6]The corpus is freely available at `http://hdl.handle.net/11356/1467`.

text was hard to annotate. For each of the two subsets the train:dev:test split is encoded in the dataset in a 60:20:20 manner.

The suitable subset has primary, secondary and tertiary labels encoded on three levels of detail – as 24 labels, 21 labels (on this level, the stratified train:dev:test split was performed) and 12 labels. Smaller sets of labels were produced by merging original categories. In this research, we perform experiments mostly on the primary labels only, to which the information from the secondary labels is added in the experiments in the subsection 4.6. We use the sets of 21 and 12 labels. The latter set is used only in the experiments in the subsection 4.7.

Each text instance is encoded as a sequence of paragraphs. In addition to the text (attribute `text`), each paragraph contains information whether the paragraph is considered a near-duplicate by automatic means (the boolean attribute `duplicate`), and finally, manually added information on whether a near-duplicate is informative for the genre identification (the boolean attribute `keep`). In this research, we exploit only the automated information of near-duplicates, as the information of near-duplicate paragraphs to be kept did not prove to be useful during our preliminary experiments. The size of the subsets is described in the Table 1.

subset	texts	pars	words
suitable	1,002	15,050	478,969
suitable (dedup.)	983	7,088	278,075
not suitable	123	3,402	173,778
both subsets	1,125	18,452	652,747

Table 1: Size of the suitable subset ("suitable" – all paragraphs, "suitable (dedup.)" – texts without near-duplicates), not suitable subset, and the sum of both, i.e., the size of the whole GINCO dataset, in terms of number of texts, paragraphs ("pars") and words.

4. Machine learning experiments

4.1. Data split

To prepare the dataset for the machine learning experiments, labels with less than 5 instances, namely *FAQ*, *Script/Drama* and *Lyrical*, were merged to the label *Other*. Thus, the experiments were performed on the set of 21 labels. The dataset was then split into train, dev and test in a 60:20:20 manner. Following Asheghi et al. (2014), we ensured that instances from the same web domain were present in only one split to minimize the effect of topic, website design, and the writing style of specific authors. Stratification by the year of crawling and the `hard` parameter was performed in a manual manner, choosing the stratification by primary label that also ensured reasonable stratification by these two additional variables.

4.2. Experimental setup

Dev split was used to optimize hyperparameters for different models. The main focus of the hyperparameter search was the number of training epochs to prevent overfitting and optimize micro and macro F1 scores. In accordance to this, 30 epochs for Transformer models and 200 epochs for fastText models were used. For the Transformer models, the sequence length of 512 tokens was used, and the learning rate was set to 10^{-5}.

The models, trained on the train split, were evaluated on the test split via micro F1 and macro F1 to measure both the instance-level and the label-level performance of a specific setup. In each experiment, at least 5 training runs were performed to assure a reasonable sample for measuring statistical significance of differences in performance of specific setups, which was tested via the Mann-Whitney U rank test.

In the remainder of this section, we present the results of our experiments:

- Section 4.3 – a comparison of different technologies at our disposal
- Section 4.4 – the impact of using full texts of web documents instead of texts with near-duplicate paragraphs removed
- Section 4.5 – investigation of the impact of the training data size
- Section 4.6 – the impact of using secondary labels as additional signal
- Section 4.7 – the impact of downcasting the number of labels from 21 to 12

4.3. Choice of technology

classifier	micro F1	macro F1
stratified dummy	0.067	0.061
fastText	0.352 ± 0.038	0.217 ± 0.040
fastText + emb.	0.361 ± 0.007	0.219 ± 0.013
XLM-RoBERTa	0.624 ± 0.015	$\mathbf{0.579 \pm 0.024}$
SloBERTa	$\mathbf{0.629 \pm 0.016}$	0.575 ± 0.037

Table 2: Comparison of classifiers. Deduplicated datasets were used for training and evaluation. 'fastText + emb' denotes fastText with pre-trained Slovenian embeddings.

To assess which technology is the most suitable for the automatic genre identification task, we compared fastText (Joulin et al., 2016) and two base-sized Transformer-based pre-trained language models – the monolingual Slovenian SloBERTa (Ulčar and Robnik-Šikonja, 2021) model and multilingual XLM-RoBERTa (Conneau et al., 2019) model. Additionally, as an illustration of the lower bound, a dummy classifier with stratified guessing strategy was implemented. The results of the experiment, summarized in

Table 2, revealed that fastText performs significantly worse than the Transformer models, and the addition of Slovenian embeddings (Ljubešić and Erjavec, 2018) only marginally increased its performance. The monolingual model SloBERTa and the multilingual model XLM-RoBERTa revealed to be the most suitable for the AGI task, with SloBERTa reaching 0.629 in micro F1 and 0.575 in macro F1, and XLM-RoBERTa 0.624 in micro F1 and 0.579 in macro F1. XLM-RoBERTa was included in the comparison following the findings of Repo et al. (2021) where it outperformed monolingual BERT models in this task. However, in our case it did not perform statistically significantly different from SloBERTa.

Two main conclusions can be drawn from these results. For identifying genre, CNN-like classifiers seem not to be up to the task, while Transformer-based models achieve a drastic improvement of the results. Furthermore, similar to results of previous research, multilingual BERT models seem to be as good for modelling the phenomenon as monolingual BERT models. We currently have two hypotheses why this is the case: 1. model pre-training might not bring a lot to the task, but rather the large Transformer model capacity, and 2. genre might be a more generic linguistic task than standard NLP tasks on which monolingual models tend to outperform multilingual models. We plan to test both of these hypotheses in our future research.

Further experiments were performed with the SloBERTa model. However, future experiments in the cross-lingual genre identification will certainly make use of the equally-performing XLM-RoBERTa model.

4.4. Impact of near-duplicate removal

A standard method for pre-processing web-based data is removal of near-duplicates. While the initial experiment presented in Section 4.3 was performed on documents with near-duplicate paragraphs removed, in these experiments we investigate whether keeping all text for classification is beneficial, especially given that for most of the categories there is not much text available.

The results of these experiments are summarized in Table 3. The main finding is that full texts tend to perform better regarding the macro F1 metric, but worse on the micro F1 metric. This result makes sense as it shows for categories with lower results to improve if more text is available per instance, but the instances from the most populous classes seem to perform worse, resulting in an overall worse per-instance performance. Given the overall minor differences in the two setups, and our greater interest in the per-instance performance, i.e., micro F1, we decided to continue our experiments on the deduplicated dataset.

dataset	micro F1	macro F1
full	0.607 ± 0.019	$\mathbf{0.596 \pm 0.033}$ *
deduplicated	$\mathbf{0.629 \pm 0.016}$ ***	0.575 ± 0.037

Table 3: Effect of near-duplicate paragraph removal. Mann-Whitney U rank test was used for p-value estimation, with the hypothesis that the distribution of one metric, marked with an asterisk, was greater than the distribution of another. Asterisks denote p-value: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

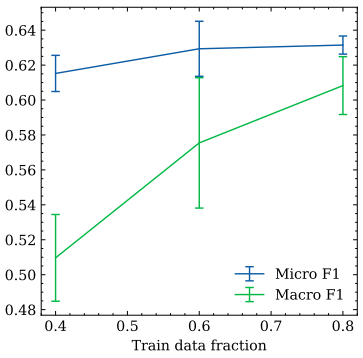


Figure 2: Effect of train data size on micro and macro F1 scores.

4.5. Impact of training data size

To understand the effect of the training data size on the performance of the models, we repeated training on train + dev, as well as on a random subset of the training data, comprising of 40% of the whole dataset. In this way, two more datapoints were obtained, with the training sizes being 40%, 60% and 80% of the entire dataset. The effect that this manipulation has on macro and micro F1 scores is shown in Figure 2. The train size correlates positively with both metrics, but the increase is lower after the second train data increase. However, further increasing the training data size should help especially with the low-frequency classes, which is supported by greater improvements obtained on macro F1 than on micro F1.

At this point we analysed the performance of the model on specific categories as well, and compared it with the frequency of each category in the dataset. Figure 3 presents F1 metrics per category, calculated when training on 60% of the data (original train split), ordered by the decreasing frequency of categories. Interestingly, the category frequency does not correlate with the F1 metric. Some of the categories, i.e., *Forum*, *Research Article*, and *Recipe*, are supported by less data than the most frequent categories, but perform better. This indicates that some categories are easier to classify than others. However, one has to bear in mind that the frequency of a category has a direct impact on the size not only of the training data, but also the test data, i.e., the F1 scores of the less represented categories depend more heavily on the (dis)similarity of the few instances in the test split to the few instances in the train split.

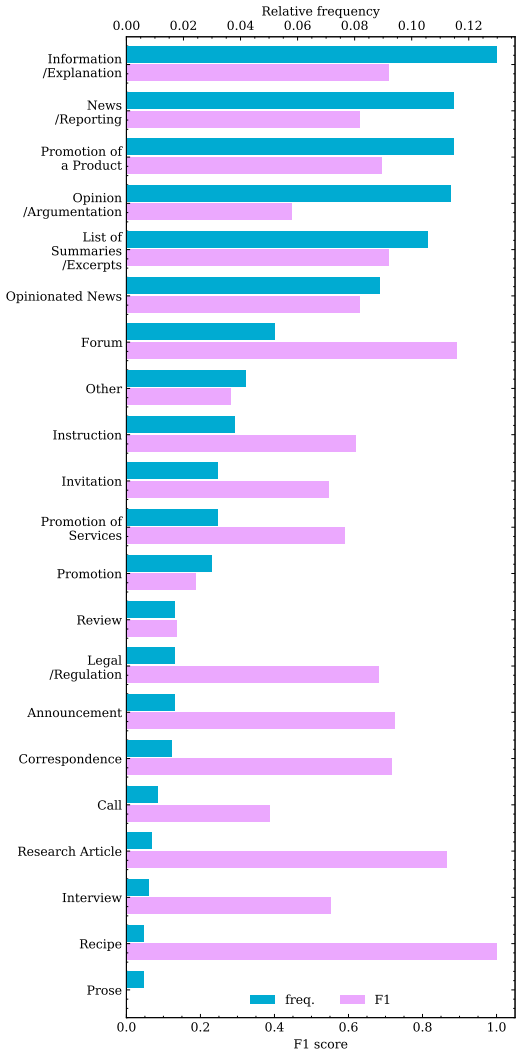


Figure 3: Per-category frequency and F1 scores for training on 60% of the deduplicated data. The blue bar depicts the relative frequency of the category, and the purple bar shows the average F1 score from all training runs for a specific category. For reading F1 scores, use bottom x axis, for relative frequencies, use top x axis.

4.6. Secondary labels as additional signal

As 188 or 18.7% of the texts are labelled with a secondary category as well, denoting presence of an additional genre, we used this information to inspect whether including the secondary labels in the train split as additional signal would improve the performance of the classifier. The experiments were performed on the original train split (60% of data) of the deduplicated dataset.

A separate training dataset was prepared, where the instances were repeated three times and the last repetition was labelled with the secondary label in an attempt to augment the performance of the models. With that, we deemed the importance of the primary label to be twice as large as the importance of the secondary label. If there was no secondary label, the instance was repeated three times with the primary label. Given that the instances were repeated three times, the number of training epochs was adapted from 30 to 10 in an attempt to make all experiments comparable. Similarly, we did not adapt the test dataset in any way. As shown in Table 4, the inclusion of secondary labels improved the micro F1 score, while the macro F1 score decreased. The increased micro F1 is not statistically significant, macro F1 decrease, however, is. To conclude, while there might be a positive impact of inclusion of the secondary label, it is minimal and it complicates the setup. Therefore, we opted for using only the primary labels in our final experiment, described in the following subsection.

train labels	micro F1	macro F1
primary	0.629 ± 0.016	$\mathbf{0.575 \pm 0.037}$ *
both	$\mathbf{0.635 \pm 0.011}$	0.558 ± 0.026

Table 4: Impact of secondary label inclusion in the training dataset. Statistical testing in each column is performed with Mann-Whitney U rank test with p-value encoded with asterisks: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

4.7. Number of classes

In the final part of experimentation, we investigate the impact of using a smaller set of labels. To this end, we reduce the number of labels from 21 to 12 labels by merging similar labels (e.g., *Promotion of a Product*, *Promotion of Services*, *Invitation* and *Promotion*) and by adding some less represented labels, such as *Prose*, to the category *Other*.

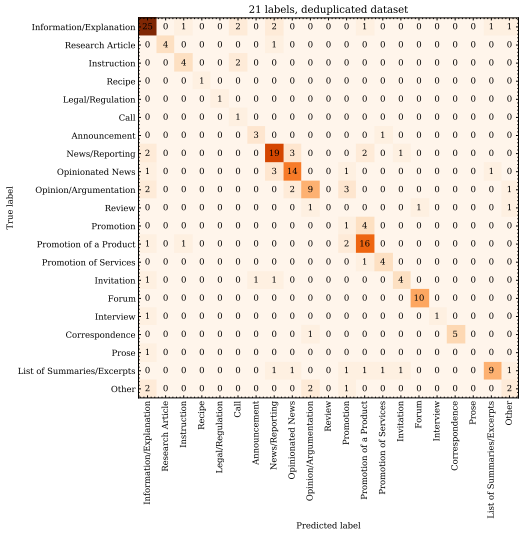
The use of a smaller label set significantly improved macro and micro F1 scores as shown in Table 5. A positive impact, especially for the categories *Promotion* and *Other*, is also visible in confusion matrices, which are presented in Figure 4a (21 labels) and Figure 4b (12 labels). Here it should be noted once more that the performance of the less represented classes, such as *Legal/Regulation*, *Call*, *Interview* and so on, heavily depends on a very small number of instances in the training and test set. In future work, we plan to double the size of the dataset to provide more training and test examples.

5. Conclusion

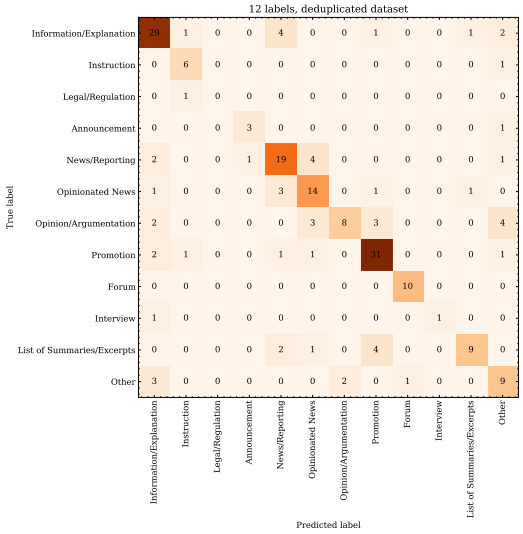
In this paper we presented a representative dataset of Slovenian web-crawled documents annotated with genre – the freely available Genre Identification Corpus GINCO 1.0 (Kuzman et al., 2021). We introduced a new genre schema which allows annotation with genre of the entire composition of not only web documents,

label set	micro F1	macro F1
21 labels	0.629 ± 0.016	0.575 ± 0.037
12 labels	$\mathbf{0.696 \pm 0.011}$ ***	$\mathbf{0.668 \pm 0.028}$ ***

Table 5: The effect of using a smaller label set by merging 21 labels to 12 labels. Deduplicated datasets were used for training and evaluation. Statistical testing in each column is performed with Mann-Whitney U rank test with p-value encoded with asterisks: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.



(a) Original label set (21 labels).



(b) Reduced label set (12 labels).

Figure 4: Confusion matrices for the best performing training run for 21 labels (4a) and 12 labels (4b).

but textual documents in general. Furthermore, we proposed some improvements in the annotation procedure with which we achieved the nominal Krippendorff's alpha (Krippendorff, 2018) of 0.71 which indicates that our approach allows more reliable genre annotation than currently most frequently used approaches.

We performed a series of experiments over the new dataset, revealing that CNN-like classifiers are not up to the task, and the language-specific SloBERTa (Ulčar and Robnik-Šikonja, 2021) Transformer model to be equally potent as the multilingual XLM-RoBERTa (Conneau et al., 2019). Furthermore, we showed that the most reasonable setup for web genre identification is to work with documents with near-duplicate paragraphs removed, using only the dominant, primary labels of texts, and that genre frequency and classification performance do not correlate. When applying this setup to experiments with SloBERTa, we reached 0.629 in micro F1 and 0.575 in macro F1.

While we performed our experiments on a random selection of the Slovenian web, we discarded 10.9% of the data as unsuitable, for which we do not have an efficient classification, or rather elimination approach, and we manually split documents, containing multiple texts with different genres, a task that we did not take on to automate. Once we have these two issues moved out of our way, we can be able to fully claim that we are able to perform high-quality genre identification on a random sample of web data. Nevertheless, this dataset and these experiments still represent the most realistic web-based sample of documents annotated for suitability and genre. Encouraged by positive results, we plan to continue with annotation campaigns to enlarge the Slovenian dataset and to create a Croatian and an English dataset, which will allow cross-lingual experiments.

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

6. Bibliographical References

Asheghi, N. R., Markert, K., and Sharoff, S. (2014). Semi-supervised graph-based genre classification for web pages. In *Proceedings of TextGraphs-9: The workshop on graph-based methods for natural language processing*, pages 39–47.

Asheghi, N. R., Sharoff, S., and Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Berninger, V. F., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the krys i corpus. In *BCS-IRSG Workshop on Corpus Profiling*, pages 1–10.

Biber, D. and Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Boese, E. S. (2005). *Stereotyping the web: genre classification of web documents*. Ph.D. thesis, Citeseer.

Bulygin, M. and Sharoff, S. (2018). Using machine translation for automatic genre classification in arabic. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 153–162.

Crowston, K., Kwaśnik, B., and Rubleske, J. (2010). Problems in the use-centered development of a taxonomy of web genres. In *Genres on the Web*, pages 69–84. Springer.

Egbert, J., Biber, D., and Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In *Proceedings of the fifth Web as Corpus workshop*, pages 27–35.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., and Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Laippala, V., Rönnqvist, S., Hellström, S., Luotolahti, J., Repo, L., Salmela, A., Skantsi, V., and Pyysalo, S. (2020). From web crawl to clean register-annotated corpora. In *Proceedings of the 12th Web as Corpus Workshop*, pages 14–22.

Laippala, V., Egbert, J., Biber, D., and Kyröläinen, A.-J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*, pages 1–32.

Laippala, V. (2019). From bits and numbers to explanations–doing research on internet-based big data. In *DATA AND HUMANITIES (RDHUM) 2019 CONFERENCE: DATA, METHODS AND TOOLS*, page 139.

Lee, Y.-B. and Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

Lee, D. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. In *Teaching and Learning by Doing Corpus Analysis*, pages 245–292. Brill Rodopi.

Lim, C. S., Lee, K. J., and Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information processing & management*, 41(5):1263–1276.

Müller-Eberstein, M., van der Goot, R., and Plank, B. (2021). Genre as weak supervision for cross-lingual dependency parsing. *arXiv preprint arXiv:2109.04733*.

Orlikowski, W. J. and Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *LREC*.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo, S., and Laippala, V. (2021). Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *arXiv preprint arXiv:2102.07396*.

Rönnqvist, S., Skantsi, V., Oinonen, M., and Laippala, V. (2021). Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., and Liu, X. (2001). Genre based navigation on the web. In *Proceedings of the 34th annual Hawaii international conference on system sciences*, pages 10–pp. IEEE.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In *Genres on the Web*, pages 3–30. Springer.

Santini, S. M. (2006). Common criteria for genre classification: Annotation and granularity. In *Workshop on Text-based Information Retrieval (TIR-06), In Conjunction with ECAI 2006, Riva del Garda, 2006*. Citeseer.

Santini, M. (2007). *Automatic identification of genre in web pages*. Ph.D. thesis, University of Brighton.

Santini, M. (2010). Cross-testing a genre classification model for the web. In *Genres on the Web*, pages 87–128. Springer.

Sharoff, S., Wu, Z., and Markert, K. (2010). The web library of babel: evaluating genre collections. In *LREC*. Citeseer.

Sharoff, S. (2010). In the garden and in the jungle. In *Genres on the Web*, pages 149–166. Springer.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*.

Stewart, J. G. and Callan, J. (2009). *Genre oriented summarization*. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. *Proc. Towards a Reference Corpus of Web Genres*.

Suchomel, V. (2020). Genre annotation of web corpora: Scheme and issues. In *Proceedings of the Future Technologies Conference*, pages 738–754. Springer.

Van der Wees, M., Bisazza, A., and Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vidulin, V., Luštrek, M., and Gams, M. (2007). Using genres to improve search engines. In *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 45–51.

Williams, Kevin Crowston, M. (2000). Reproduced and emergent genres of communication on the world wide web. *The information society*, 16(3):201–215.

Zu Eissen, S. M. and Stein, B. (2004). Genre classification of web pages. In *Annual Conference on Artificial Intelligence*, pages 256–269. Springer.

7. Language Resource References

Bañón, Marta and Esplà-Gomis, Miquel and Forcada, Mikel L. and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and van der Werff, Tobias and Zaragoza, Jaume. (2022). *Slovene web corpus MaCoCu-sl 1.0*.

Alexis Conneau and Kartikay Khandelwal and Naman Goyal and Vishrav Chaudhary and Guillaume Wenzek and Francisco Guzmán and Edouard Grave and Myle Ott and Luke Zettlemoyer and Veselin Stoyanov. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*.

Erjavec, Tomaž and Ljubešić, Nikola. (2014). *The slWaC 2.0 corpus of the Slovene web*.

Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas. (2016). *Bag of tricks for efficient text classification*.

Kuzman, Taja and Brglez, Mojca and Rupnik, Peter and Ljubešić, Nikola. (2021). *Slovene Web genre identification corpus GINCO 1.0*.

Ljubešić, Nikola and Erjavec, Tomaž. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*.

Pomikálek, Jan. (2011). *Removing boilerplate and duplicate content from web corpora*.

Suchomel, Vít and Pomikálek, Jan. (2012). *Efficient Web Crawling for Large Text Corpora*.

Ulčar, Matej and Robnik-Šikonja, Marko. (2021). *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*.

Appendix 1: Genre Categories

Genre	Description
News/Reporting	an objective text which reports on an event recent at the time of writing or coming in the near future
Announcement	an objective text which notifies the readers about new circumstances, asking them to act accordingly
Instruction	an objective text which instructs the readers on how to do something
Recipe	an objective text which instructs the readers on how to prepare food or drinks
Information/Explanation	an objective text that describes or presents a person, a thing, a concept etc.
Research Article	an objective text which presents research, uses formal language and scientific terms
Call	a text which asks the readers to submit a paper, project proposal, original literary text etc., stating requirements and a deadline
Legal/Regulation	an objective formal text that contains legal terms and is clearly structured
Opinionated News	a subjective text which reports on an event recent at the time of writing or coming in the near future
Opinion/Argumentation	a subjective text in which the authors convey their opinion or narrate their experience. It includes promotion of an ideology and other non-commercial causes.
Review	a subjective text in which authors evaluate a certain entity based on their personal experience
Promotion	a subjective text intended to sell or promote something. This category is used when a promotional text does not fall under Promotion of a Product, Promotion of Services or Invitation.
Promotion of a Product	a subjective text which promotes a product, an application, an accommodation, etc.
Promotion of Services	a subjective text which promotes services of a company
Invitation	a text which invites the readers to participate in an event
Interview	a text consisting of questions posed by the interviewer and answers by the interviewee
Forum	a text in which people discuss a certain topic in form of comments
Correspondence	a text addressed to a person or organization with a form similar to a letter, i.e., including a greeting, a complimentary close etc.
Script/Drama	a literary text that mostly consists of dialogue of characters, stage directions and instructions to the actors
Lyrical	a text that consists of verses
Prose	a literary running text that consists of paragraphs
FAQ	a text in which an author informs the reader through questions and answers
List of Summaries/Excerpts	a text which consists of summaries or excerpts of multiple articles/topics (usually from the article archive page)
Other	a text that has a purpose, not covered by other genre categories, or has no clear purpose

Table 6: Description of genre categories

3.2 Merging Multiple High-Coverage Genre Schemata and Datasets

While Transformer-based models have shown promising performance on the Slovenian genre dataset (see Section 3.1.4), the results still fall short of our target of 0.80 in both micro-F1 and macro-F1 scores. Previous studies have shown that fine-tuning on multiple datasets can improve performance, while also assuring a higher robustness of the model by mitigating topical biases (Lepekhin & Sharoff, 2022). Therefore, in this section, we construct a new Slovenian-English dataset by merging three manually-annotated genre datasets in two languages: the English CORE (Egbert et al., 2015) and the FTD dataset (Sharoff, 2018), and the Slovenian GINCO dataset (see Section 3.1.3).

The first step in this endeavor involves assessing the feasibility of conducting cross-dataset experiments between the GINCO and CORE (Egbert et al., 2015) datasets. Our findings show that when the XLM-RoBERTa model (Conneau et al., 2020) is fine-tuned on one dataset and evaluated on the other, the scores remain relatively high – ranging from 0.60 to 0.64 in micro-F1 and 0.54 to 0.58 in macro-F1 scores. This confirms that the GINCO and CORE datasets are sufficiently comparable to allow cross-dataset transfer. Moreover, these results are comparable to those from cross-lingual experiments using English, Finnish, Swedish, and French datasets based on the same CORE schema (Repo et al., 2021). For a more detailed analysis, see the paper "*Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments*" by Kuzman, Ljubešić, and Pollak (2022), enclosed in Section 3.2.3.

Motivated by the positive results, we extend the unified schema to incorporate the FTD categories (Sharoff, 2018), resulting in the X-GENRE schema (see Section 3.2.1). Based on this new schema, we combine the GINCO, CORE, and FTD datasets to create the Slovenian–English X-GENRE dataset, presented in Section 3.2.2. With this novel dataset, we address Goal *G3* by providing a manually-annotated genre dataset in both Slovenian and English. The X-GENRE dataset offers greater robustness through its multilingual content and a broader variety of training and test instances. The remainder of this section is based on the paper "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" by Kuzman, Mozetič, et al. (2023), enclosed in Section 3.2.4.

3.2.1 X-GENRE Schema

Each of the CORE (Egbert et al., 2015), FTD (Sharoff, 2018), and GINCO datasets uses its own genre schema with varying label granularity. To merge them, we develop a unified genre schema – referred to as the X-GENRE schema – to which the original labels are mapped. The development of this schema involves analyzing label descriptions, reviewing manual annotation guidelines, and examining concrete examples from all three datasets. In addition, we adopt a data-driven approach: Transformer-based BERT-like models are fine-tuned on each dataset using its original schema, and then applied to the other two datasets. This helps us identify which labels truly match – not just in theory, but also based on the predictions of the models across datasets. Some labels were found to be spread across multiple categories and could not be reliably mapped to a single label from the other two schemata. These labels were therefore discarded, and texts annotated with them were excluded from the final X-GENRE dataset. The mapping of the GINCO, FTD, and CORE schemata to the unified X-GENRE schema is shown in Figure 3.4.

The resulting X-GENRE schema comprises 9 genre labels: *Information/Explanation, Instruction, Legal, News, Opinion/Argumentation, Promotion, Prose/Lyrical, Forum*, and *Other*. Label descriptions are provided in the Appendix of Kuzman, Mozetič, et al. (2023)

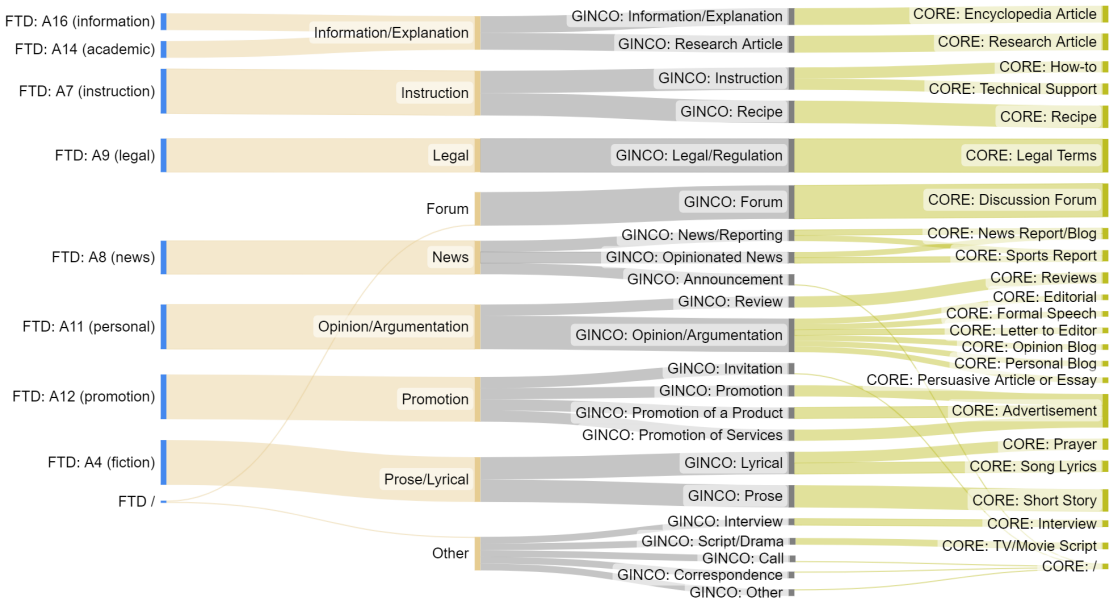


Figure 3.4: The mapping of genre labels from the FTD (Sharoff, 2018) (first column), GINCO (Kuzman, Rupnik, et al., 2022) (third column), and CORE (Egbert et al., 2015) (fourth column) schemata to the unified X-GENRE schema (second column).

(see Section 3.2.4), and further details, including examples and a decision tree, are available in the publicly accessible annotation guidelines.[3]

3.2.2 Slovenian-English X-GENRE Dataset

The unified X-GENRE schema allows us to merge parts of the English CORE (Egbert et al., 2015) and FTD (Sharoff, 2018) datasets with the Slovenian GINCO dataset. The resulting X-GENRE dataset (Kuzman & Ljubešić, 2024a) contains 2,956 instances, with roughly equal contributions from each source, namely, around 1,000 instances per dataset. It is split into training, evaluation, and test sets following a 60:20:20 ratio (1,772:592:592 texts), with a similar distribution of genre labels and dataset sources throughout the splits to ensure a balanced representation of English and Slovenian texts. The X-GENRE dataset is available on both the CLARIN.SI and Hugging Face repositories to ensure long-term preservation and open access.[4]

The X-GENRE dataset fulfills the part of Goal *G3* related to constructing Slovenian and English training genre datasets. The merged dataset offers several advantages. It brings together three datasets in two languages, collected from different sources, time periods, and using different methods. This diversity provides a wider range of instances, enhancing the generalization and robustness of models trained on it. Additionally, the inclusion of both English and Slovenian languages allows analyses of the multilingual performance of genre classifiers.

To assess the suitability of the newly developed dataset for training genre classifiers,

[3]The annotation guidelines for the X-GENRE schema are available at https://tajakuzman.github.io/X-GENRE-Annotation-Guidelines/.

[4]The X-GENRE dataset is accessible on the CLARIN.SI repository at http://hdl.handle.net/11356/1960 and the Hugging Face repository at https://doi.org/10.57967/hf/6582.

compared to existing high-coverage genre datasets, we fine-tune and evaluate the base-sized XLM-RoBERTa Transformer model (Conneau et al., 2020) on each dataset. The results, presented in Table 3.3, show that the model fine-tuned on the X-GENRE dataset achieves the highest results, namely, micro-F1 of 0.80 and macro-F1 of 0.79. This finding confirms that training on multiple datasets improves performance, which is consistent with the earlier results of Repo et al. (2021) and Lepekhin and Sharoff (2022). Moreover, the X-GENRE dataset has the same number of labels as the CORE dataset when considering its main categories, yet it is six times smaller. Despite this, the models fine-tuned on the X-GENRE dataset achieve significantly better results, demonstrating that this dataset is more effective for training genre classifiers and reinforcing the notion that dataset quality is more important than size. Full details on the methodology and experimental setup are available in the paper by Kuzman, Mozetič, et al. (2023), enclosed in Section 3.2.4.

Table 3.3: Performance of XLM-RoBERTa-based genre classifiers, fine-tuned and evaluated on GINCO, CORE (Egbert et al., 2015), FTD (Sharoff, 2018), and X-GENRE datasets and their corresponding schemata. The results are ordered based on the macro-F1 scores.

Dataset	Labels	Training Instances	Micro-F1	Macro-F1
X-GENRE	9	1,772	**0.80**	**0.79**
FTD	10	849	0.74	0.74
GINCO (simplified schema)	9	601	0.73	0.72
CORE (main categories)	9	10,256	0.75	0.62
GINCO (reduced schema)	17	579	0.59	0.47
CORE (subcategories)	37	9,537	0.66	0.39

Since the results show that the X-GENRE dataset is the most suitable for training genre classifiers, we use this dataset in our comparison of different machine learning models in in-domain and out-of-domain scenarios, presented in the following chapter.

3.2.3 Related Paper: Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Taja Kuzman[†] [*], Nikola Ljubešić[†], Senja Pollak[†]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
taja.kuzman@ijs.si, nikola.ljubesic@ijs.si, senja.pollak@ijs.si

[*]Jožef Stefan International Postgraduate School

Abstract

This article explores comparability of an English and a Slovene genre-annotated dataset via monolingual and cross-lingual experiments, performed with two Transformer models. In addition, we analyze whether translating the Slovene dataset into English with a machine translation system improves monolingual and cross-lingual performance. Results show that cross-lingual transfer is possible despite the differences between the datasets in terms of genre schemata and corpora construction methods. Furthermore, the XLM-RoBERTa model was shown to provide good results in both settings already when learning on less than 1,000 instances. In contrast, the trilingual CroSloEngual BERT model was revealed to be less suitable for this text classification task. Moreover, the results reveal that although the English dataset is 40 times larger than the Slovene dataset, it provides similar or worse classification results.

1. Introduction

Texts in datasets can be grouped by genres based on their common function, form and the author's purpose (Orlikowski and Yates, 1994). Labeling texts with genres allows for a deeper insight into the composition and quality of a web corpus that was collected with automatic means, more efficient queries in information retrieval tools (Vidulin et al., 2007), as well as improvements of various language technologies tasks, such as part-of-speech tagging (Giesbrecht and Evert, 2009) and machine translation (Van der Wees et al., 2018). That is why automatic genre identification (AGI) has been a subject of numerous studies in the computational linguistics and information retrieval fields (e.g., see Egbert et al. (2015), Sharoff (2018)).

As in other text classification tasks, a large manually annotated dataset is required in AGI in order to train and test a classifier. While there exist some large English genre-annotated datasets, such as the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) with 53,000 texts and the Leeds Web Genre Corpus (Asheghi et al., 2016) with 5,000 texts, for other languages there is either no dataset or mostly a small one, consisting of 1,000 to 2,000 texts, such as genre-annotated corpora for Russian (Sharoff, 2018), Finnish (Laippala et al., 2019), Swedish and French (Repo et al., 2021). This means that for obtaining a large dataset needed for genre identification of other languages, costly and time-consuming annotation campaigns are still needed, leaving most languages under-resourced in regard to the technologies based on the AGI.

However, it might be possible to overcome this obstacle by leveraging the cross-lingual transfer, applying models trained on high-resource languages to the low-resource languages. Recently, Repo et al. (2021) showed that it is possible to achieve good levels of cross-lingual transfer in AGI experiments. They performed experiments in zero-shot cross-lingual automatic genre identification by training multilingual Transformer-based models on the English CORE corpus (Egbert et al., 2015) and testing them on smaller Finnish, Swedish, and French datasets. Rönnqvist et al. (2021) extended this research, training the models on a multilingual dataset, created from the four corpora, which further improved the results.

These promising results stimulated creation of genre-annotated datasets for other languages, and for Slovene, a web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) was created. Its genre schema was based on the CORE schema with the possibility of cross-lingual experiments in mind (see Kuzman et al. (2022)). However, a linguistic analysis of the categories (Biber and Egbert, 2018) and a low inter-annotator agreement, reported by Egbert et al. (2015) and Sharoff (2018), revealed some shortcomings of the CORE schema that could impact the reliability of the dataset. Thus, Kuzman et al. (2022) diverged from the original schema when annotating GINCO, striving towards a more reliably annotated dataset. In addition to this, the CORE and GINCO datasets were created following different corpora collection and annotation approaches (see Section 3.1.). Due to these differences, it remained unclear whether the datasets are comparable enough to allow cross-lingual transfer which would eliminate the need for extensive annotation campaigns of Slovene and other under-resourced languages of interest. This article provides first insight into this, exploring the comparability of the two datasets through cross-dataset and cross-lingual experiments.

2. Goal of the Paper

This paper analyzes comparability of two genre-annotated datasets, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021). We perform cross-dataset and cross-lingual automatic genre identification experiments to address the main research question (Q1): Is the CORE dataset comparable to the GINCO dataset enough to provide good cross-lingual transfer, as it was achieved by Repo et al. (2021) who

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

used comparably encoded Finnish, Swedish and French datasets?

To compare the corpora and to analyze their usefulness for monolingual as well as for cross-lingual automatic genre identification, first, labels from both corpora were mapped to a joint schema, the GINCORE schema. Then, multilingual pre-trained Transformer-based models were trained on the English CORE dataset with GINCORE labels (EN-GINCORE), the Slovene GINCO dataset with GINCORE labels (SL-GINCORE) and the SL-GINCORE dataset that was machine translated into English (MT-GINCORE). We conduct 1) monolingual in-dataset AGI experiments, training and testing on the same dataset, 2) cross-lingual and cross-dataset AGI experiments, training on one dataset and testing on the other. The machine-translated dataset is added to the comparison to explore two additional research questions: Q2) In monolingual in-dataset experiments, do multilingual models, which were pre-trained on more English than Slovene data, perform differently on Slovene dataset (SL-GINCORE) than on a Slovene dataset, machine-translated to English (MT-GINCORE)? and Q3) In cross-lingual cross-dataset experiments, does translating the training data (MT-GINCORE) into the language of test data (EN-GINCORE) provide better results than using training and testing data in different languages (SL-GINCORE and EN-GINCORE)?

The experiments were performed with two multilingual Transformer-based pre-trained language models, massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). This provides an answer to the fourth research question (Q4): Does CroSloEngual BERT, pre-trained on a smaller number of languages, perform better in the cross-lingual AGI experiments than a massively multilingual XLM-RoBERTa model?

3. Data Preparation

3.1. Original Datasets

In this research, three datasets were used: the Corpus of Online Registers of English (CORE) (Egbert et al., 2015), the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) and the GINCO 1.0 corpus, machine translated to English.

The CORE corpus consists of web texts that were extracted from the "General" part of the Corpus of Global Web-based English (GloWbE) (Davies and Fuchs, 2015). The GloWbE corpus was collected via Google searches with high frequency English 3-grams as the queries (Davies and Fuchs, 2015). After obtaining the texts, further cleaning was performed, more specifically, the boilerplate was removed with the Justext tool (Pomikálek, 2011).

The CORE corpus was annotated based on a hierarchical schema which consists of 8 main genre categories, such as *Narrative*, *Opinion*, *Spoken*, and 54 subcategories, e.g., *News Report/Blog*, *Instruction*, *Travel Blog*, *Magazine Article*. The annotation was single-label, i.e., each annotator, recruited through a crowd-sourcing platform, could assign one main category and one subcategory to a text. However, as each text was annotated by four annotators, that means that it can have up to four labels. The corpus that we obtained from the authors and used in this research consists of 48,415 texts, labeled with 8 main categories and 47 subcategories. The corpus was further cleaned by removing duplicated texts and texts with more than one assigned label, resulting in 41,502 texts.

The GINCO corpus (Kuzman et al., 2022) consists of a random sample of web texts from two Slovene web corpora, slWaC 2.0 corpus (Erjavec and Ljubešić, 2014) from 2014 and MaCoCu-sl 1.0 corpus (Bañón et al., 2022) from 2021. Both web corpora were created by crawling the Slovene top-level domain and some generic domains that are interlinked with the national domain. As in GloWbE, the boilerplate was removed with the Justext tool (Pomikálek, 2011). The GINCO corpus consists of two parts, the "suitable" part, annotated with genres, and "not suitable" part, consisting of texts not suitable for genre annotation, such as texts in other languages, machine-translated texts etc. In this research, only the suitable part, consisting of 1,002 texts, was used.

For the annotation, a GINCO schema was used, consisting of 24 labels, e.g., *News/Reporting*, *Opinion/Argumentation*, *Promotion of a Product*. The schema is based on the subcategory level of the CORE schema and on other schemata from previous genre studies. The texts were annotated by two annotators with the background in linguistics. In case of disagreement, final labels were determined at frequent meetings. Multi-label annotation was allowed, i.e., each text could be annotated with up to three classes which were ordered according to their prevalence in the text as a primary, secondary and tertiary label. However, in these experiments, only the primary labels are used. Each paragraph in the texts is accompanied with metadata (attribute `keep`) with information on whether it was manually identified to be a part of the main text and thus useful for the annotation. In this research, paragraphs not deemed to be useful were discarded.

The machine-translated GINCO corpus (MT-GINCO) was created by translating the Slovene GINCO 1.0 to English with the DeepL[1] machine translation system. The system is stated by its developers to be "3x more accurate" than its closest competitors, i.e., Google Translate, Amazon Translate and Microsoft Translator, based on internal blind tests (DeepL, nd). DeepL was confirmed to outperform Google Translate also in an independent study of Yulianto and Supriatnaningsih (2021). The GINCO corpus was translated into British English, as this variety seems to be more frequent than American English in the general part of the GloWbE corpus on which the CORE corpus is based (GloWbe, nd). The prevalence of the British variety in the CORE corpus was also confirmed with a lexicon-based American-British-variety Classifier (Rupnik et al., 2022) which identified 40% of texts to be British, 25% to be American, while the rest contain a mixture of both varieties or no signal for any of them.

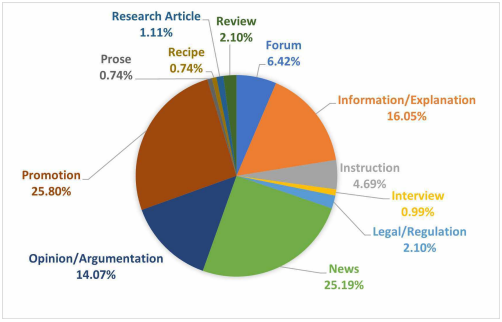
3.2. GINCORE Schema

To be able to perform cross-dataset experiments, the CORE and GINCO schemata were mapped to a joint

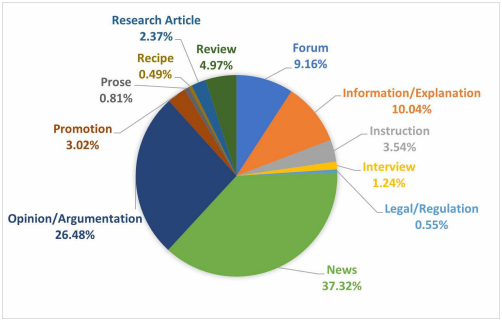
[1] `https://www.deepl.com/translator`

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022



(a) SL-GINCORE and MT-GINCORE.



(b) EN-GINCORE.

Figure 1: The differences between the distributions of GINCORE labels in the GINCO corpora MT-GINCORE and SL-GINCORE, and in the EN-GINCORE (CORE corpus).

schema – the GINCORE schema. The schemata were mapped based on descriptions of categories in previous research, in the annotation guidelines for GINCO[2] and the guidelines for CORE, created for the needs of annotation of Finnish, French and Swedish corpora using the CORE schema[3] in further research (Laippala et al., 2019; Laippala et al., 2020). Furthermore, manual inspection of instances from the GINCO and CORE corpora was performed to analyze to which extent the annotations in the corpora match the guidelines. The basis of the GINCORE schema was the GINCO schema as it was shown to provide a more reliable annotation than CORE (see Kuzman et al. (2022)). Moreover, it is easier to map 54 CORE subcategories with a very high granularity to 24 broader GINCO categories than vice versa. The CORE schema consists of broad main categories and more specific subcategories. As the GINCO schema was based on the subcategories of the CORE schema, the subcategories level was used for the mapping from CORE to GINCORE.

Some of genre categories in both schemata are identical and can be directly mapped, namely *Recipe*, *Review*, *Interview* and *Legal/Regulation*. As the GINCO and CORE schemata differ in granularity, broader GINCORE labels were created which efficiently cover categories from both schemata. Some CORE categories were not included in the mapping, because a) these labels revealed to be very infrequent and there is no sufficient information about them available, b) the labels were too broad or problematic for annotators and as a result include instances that are too heterogeneous and cannot be mapped to just one GINCORE label. The resulting GINCORE schema[4] covers 43 CORE subcategories and all 24 GINCO categories by using 20 labels: 15 labels that are present in both corpora, and 5 labels, newly introduced by the GINCO schema and thus present only in the GINCO dataset.

3.3. GINCORE Datasets

For the purpose of performing cross-dataset experiments, only the GINCORE classes that have more than 5 instances in each of the datasets were used, resulting in a smaller set of 12 GINCORE labels: *News*, *Forum*, *Opinion/Argumentation*, *Review*, *Research Article*, *Information/Explanation*, *Promotion*, *Instruction*, *Prose*, *Interview*, *Legal/Regulation*, and *Recipe*. The texts annotated with other GINCORE labels were not included in the experiments. Thus, the final datasets are slightly smaller:

- the English CORE dataset with 12 GINCORE labels, henceforth referred to as the English GINCORE dataset (EN-GINCORE), consists of 33,918 texts;
- the Slovene GINCO dataset with 12 GINCORE labels, henceforth referred to as the Slovene GINCORE dataset (SL-GINCORE), consists of 810 texts;
- the machine-translated English GINCO dataset with 12 GINCORE labels, henceforth referred to as the Machine-Translated GINCORE dataset (MT-GINCORE), consists of 810 texts.

The text instances were not pre-processed, i.e. each instance is a running text as it was extracted from the original web page from which the boilerplate and HTML tags were removed. In GINCO datasets (SL-GINCORE and MT-GINCORE), the texts consist of paragraphs, which is indicated by the <p> tag, while in the CORE dataset (EN-GINCORE), the partitioning into paragraphs is not preserved. In addition to this, the datasets differ significantly in terms of length of the texts. In the CORE dataset, the median length is 649 words, while the minimum and maximum text length is 52 words and 118,278 words respectively. In the GINCO datasets, most texts are significantly shorter, with the median length of 198 words, minimum length of 12 words and maximum length of 4,134 words. As the Transformer models, used in the experiments, can

[2]The guidelines for GINCO are available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

[3]The guidelines for the annotation campaigns using the CORE schema are available here: https://turkunlp.org/register-annotation-docs/.

[4]The final table with all the GINCORE mappings is available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/genre_pages/GINCORE_mapping.html.

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

process maximum instance length of 512 tokens, this means that while the models will in most cases be trained on complete texts from the GINCO datasets, more than half of the texts from the CORE dataset will not be used in their entirety and the models will be trained only on the first part of these instances.

Here, it should be also noted that the CORE dataset and the GINCO datasets are characterized by a different distribution of GINCORE classes. Frequency of some classes, such as *Promotion*, is significantly different, as can be seen in Figure 1.

4. Machine Learning Experiments

4.1. Models

Experiments were performed with the Transformer-based pre-trained language models which were shown to perform well in the automatic genre identification task in a monolingual as well as a cross-lingual setting (Repo et al., 2021). More specifically, two models were used, the base-sized massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). The XLM-RoBERTa model was chosen because it was revealed to be the best performing model in cross-lingual automatic genre identification based on the CORE dataset (Repo et al., 2021), and to be comparable to the Slovene monolingual model SloBERTa (Ulčar and Robnik-Šikonja, 2021) in experiments, performed on GINCO (Kuzman et al., 2022). The CroSloEngual BERT model was revealed to achieve results comparable to the XLM-RoBERTa model or to even outperform the latter model in common monolingual and cross-lingual NLP tasks (Ulčar et al., 2021). Thus, it was included in these experiments to explore whether it achieves similar results on the AGI task as well.

4.2. Experimental Setup

The datasets were split into 60:20:20 train, dev and test splits, stratified according to the label distribution. The models were trained on the train split, consisting of 20,350 texts in the case of EN-GINCORE, and of 486 texts in the case of SL-GINCORE and MT-GINCORE, and tested on the test split, i.e., 6,784 texts in the case of EN-GINCORE and 162 texts in the case of SL-GINCORE and MT-GINCORE. The dev split, which is of the same size as the test split, was used for testing the hyperparameter optimization. When splitting the datasets, it was assured that the splits of SL-GINCORE and MT-GINCORE contain the same instances, so that they differ only in the language of the content.

The Transformer models are available at the Hugging Face repository and were trained using the Simple Transformers library. To find the optimal number of epochs and the learning rate, the hyperparameter search was performed separately for CroSloEngual BERT and XLM-RoBERTa. The maximum sequence length was set to 512 tokens and other hyperparameters were set to default values. As the EN-GINCORE dataset is more than 40 times larger than the SL-GINCORE and MT-GINCORE datasets, separate hyperparameter searches for each dataset were performed. Optimum learning rate was revealed to be 10^{-5}, while the optimum number of epochs varies based on the training dataset and the model, i.e., the optimum number of epochs when training on the EN-GINCORE with a) XLM-RoBERTa is 9, and b) CroSloEngual BERT is 6; while the optimum number of epochs when training on the SL-GINCORE and MT-GINCORE with a) XLM-RoBERTa is 60, and b) CroSloEngual BERT is 90.

We performed monolingual in-dataset experiments and cross-lingual cross-dataset experiments[5]. The monolingual experiments, described in Section 4.3.1., are in-dataset experiments, which means that the models were trained and tested on splits from the same dataset. In contrast to this, in cross-dataset experiments, presented in Section 4.3.2., the models are trained on one dataset and tested on the other. At the same time, these experiments are cross-lingual, as the original datasets are in different languages.

Three runs of each experiment were performed and average results are reported. The models used in monolingual and cross-lingual setups were evaluated via micro F1 and macro F1 scores to measure the instance-level and the label-level performance.

4.3. Results

4.3.1. Monolingual In-dataset Experiments

First, the datasets are compared via monolingual in-dataset experiments where the models were trained and tested on the splits of the same dataset. In addition to this, a dummy classifier which predicts the majority class was implemented as an illustration of the lower bound. The results, presented in Table 1, show that the mapping of the original labels into a joint schema was successful and that it is possible to achieve good results when learning Transformer models on GINCORE datasets. Transformer models are shown to be very effective at this task, achieving micro and macro F1 scores that are higher than the scores of the dummy model for at least 30 points. XLM-RoBERTa, which was revealed to be the best performing model, achieved relatively high results, with micro and macro F1 scores ranging between 0.72 and 0.84, even when trained on the two smaller datasets, which consist of less than 1,000 instances.

The results show that in a monolingual setting, the massively multilingual XLM-RoBERTa model outperforms the trilingual CroSloEngual BERT model. While Ulčar et al. (2021) showed that the trilingual model is comparable to the XLM-RoBERTa model at NLP tasks which are focused on classification of words or multiword units, such as named-entity recognition and part-of-speech tagging, these results reveal that CroSloEngual BERT is not as suitable as XLM-RoBERTa for automatic genre identification.

Among all monolingual experiments, the best micro and macro F1 results were achieved when the XLM-RoBERTa was trained and tested on the machine-translated MT-GINCO dataset, reaching average micro and macro F1 scores of 0.81 and 0.84 respectively. At the same time, the

[5]The code for data preparation and machine learning experiments is available here: `https://github.com/TajaKuzman/Cross-Lingual-and-Cross-Dataset-Experiments-with-Genre-Datasets`.

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

Datasets		Majority classifier		XLM-RoBERTa		CroSloEngual BERT	
Trained on	Tested on	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
SL-GINCORE	SL-GINCORE	0.259	0.027	0.782±0.02	0.725±0.01	0.738±0.01	0.599±0.06
MT-GINCORE	MT-GINCORE	0.259	0.027	**0.807**±0.01	**0.841**±0.03	0.714±0.00	0.501±0.05
EN-GINCORE	EN-GINCORE	0.363	0.036	0.768±0.00	0.715±0.00	**0.761**±0.00	**0.706**±0.00
SL-GINCORE	EN-GINCORE	0.029	0.004	**0.639**±0.01	0.539±0.01	0.547±0.02	0.391±0.02
MT-GINCORE	EN-GINCORE	0.029	0.004	0.625±0.01	0.521±0.01	0.585±0.01	0.409±0.01
EN-GINCORE	SL-GINCORE	0.253	0.027	0.603±0.02	0.575±0.03	0.566±0.02	0.510±0.03
EN-GINCORE	MT-GINCORE	0.253	0.027	0.630±0.02	**0.663**±0.03	**0.630**±0.01	**0.543**±0.01

Table 1: Results of monolingual and cross-lingual experiments performed with XLM-RoBERTa and CroSloEngual BERT models, reported via micro and macro F1 scores (averaged over three runs). As a baseline, the scores of a majority classifier are added. The best scores for each of the two Transformer models for each of the two setups (in-dataset experiments and cross-dataset experiments) are shown in bold.

lowest scores, i.e., micro F1 of 0.71 and macro F1 of 0.50, were obtained on the same dataset in combination with the CroSloEngual BERT. Similarly, while XLM-RoBERTa achieved the worst results when trained and tested on the EN-GINCORE, CroSloEngual BERT achieved the best results on this dataset. The difference between the results on the same datasets shows the importance of analyzing the output of multiple models before reaching any conclusion regarding the datasets – if only XLM-RoBERTa would be used, one could assume that the EN-GINCORE dataset is less suitable for automatic genre identification experiments. However, after performing experiments with both models, we can see that no dataset consistently provides the best results.

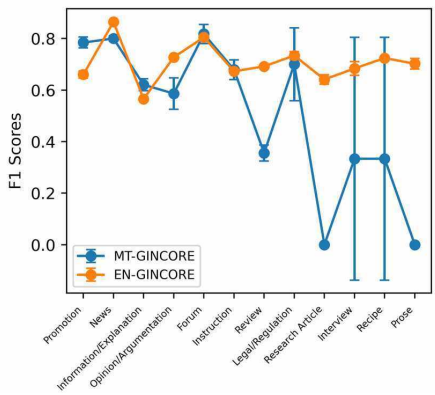


Figure 2: F1 scores per labels (averaged over three runs) in in-dataset experiments with MT-GINCORE and EN-GINCORE, performed with CroSloEngual BERT. Labels are ordered according to their frequency in the smallest of the two datasets, MT-GINCORE.

If we compare experiments, performed with the same model, we can observe that the largest differences between the datasets are in terms of macro F1 scores which are calculated on the level of labels. As shown in Figure 2, the biggest differences between the F1 scores per labels occur in cases of labels that are represented by a very small number of instances in the smaller datasets, SL-GINCORE and MT-GINCORE. Half of the labels, i.e., *Review*, *Legal/Regulation*, *Research Article*, *Interview*, *Recipe* and *Prose*, are represented by solely 4 instances or less in SL-GINCORE and MT-GINCORE test splits. One should be aware that this means that a correct or incorrect prediction of such a small number of instances per labels has a large impact on the macro F1 score. Furthermore, a correct prediction of labels with only one or two instances in the test split might happen due to chance or a similarity of texts in the train and test split. Thus, the F1 scores of these labels are not reliable. As shown in Figure 2, in the three runs, the F1 scores of *Interview* and *Recipe*, which are represented by only 1 instance in the SL-GINCORE and MT-GINCORE test sets, were either 0 or 1, which has a large impact on a macro F1. These results also show how important it is to repeat each experiment multiple times, to ascertain stability and reliability of results.

If we compare the three datasets based on micro F1 scores, there are small differences between them, i.e., a difference of 4 points between the lowest and highest scores when XLM-RoBERTa was used and a difference of 5 points when CroSloEngual BERT was used. Interestingly, although the EN-GINCORE is 40 times larger than the SL-GINCORE and MT-GINCORE, it does not provide higher results than the other two datasets when the XLM-RoBERTa model is used for training. Similar results were revealed in previous work (see Repo et al. (2021)) where they performed monolingual experiments with XLM-RoBERTa on the CORE dataset and three smaller genre-annotated datasets, Finnish FinCORE, French FreCORE and Swedish SweCORE datasets. Although the non-English datasets were annotated with the CORE schema, the annotation procedure and dataset collection methods are more similar to the GINCO approach than CORE. Their experiments showed that the XLM-RoBERTa and other Transformer models perform similarly or better when trained on datasets which consisted of 1,800 to 2,200 instances than when trained on the CORE dataset.

We have two hypotheses why this is the case: 1) It might be that due to high capacities of Transformer models, their performance on this task plateaus already at a few thousand instances and contributing bigger datasets does not significantly improve the results. 2) Or this could indicate that

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

the CORE dataset is less suitable for AGI machine learning experiments. The reason for that could be that as crowdsourcing was used for the annotation of the dataset, the assigned labels are less reliable and the classes are consequently fuzzier. Poor reliability of the dataset was also confirmed by low inter-annotator agreement. The authors of the dataset reported that there was no agreement between at least three of four annotators on the subcategory of 48.98% of texts (Egbert et al., 2015). When the schema and approach was used by Sharoff (2018) on another corpus, he reported nominal Krippendorff's alpha of 0.53 on the level of subcategories, which is below the acceptable threshold of 0.67, as defined by Krippendorff (2018). In contrast to this, the GINCO dataset was reported to achieve Krippendorff's alpha of 0.71, confirming much higher reliability of annotations.

4.3.2. Cross-lingual Cross-dataset Experiments

To assess comparability of the English CORE dataset and the Slovene GINCO dataset, we performed cross-lingual cross-dataset experiments by training the Transformer models on one dataset and testing them on another. In addition to experimenting with cross-lingual transfer from Slovene to English dataset and vice versa, we also explored whether translating the Slovene dataset into English with a machine translation system improves the results of cross-dataset experiments.

The results, shown in Table 1, reveal that the trilingual CroSloEngual BERT model performs worse than the massively multilingual XLM-RoBERTa model in the cross-lingual experiments with a difference of 12 points between the highest macro F1 scores obtained by the models and a much slighter difference between the highest micro F1 scores (0.009).

In general, results obtained in the cross-lingual experiments are significantly lower than the results from the monolingual experiments. If we compare experiments performed with XLM-RoBERTa, there are differences in 13–18 points in micro F1 and 5–32 points in macro F1 between testing the model on the same dataset as it was trained on (monolingual experiments) and on another dataset (cross-lingual experiments). In case of CroSloEngual BERT, the differences between testing on the same dataset versus testing on the other dataset were in 13–20 points in micro F1 and 9–20 points in macro F1.

Nevertheless, the XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 for micro and macro F1 respectively, are a promising indicator that cross-lingual transfer could be possible in this task for Slovene as well. Furthermore, the results are comparable to the results of cross-lingual experiments with the CORE corpora, reported by Repo et al. (2021). When they trained the XLM-RoBERTa model on the CORE corpus and tested it on Finnish, Swedish and French datasets, annotated with the CORE schema, the micro F1 scores ranged from 0.61 to 0.69. Here it needs to be noted that they used a large-sized model which was shown to significantly outperform the base-sized model used by us (Conneau et al., 2020), and that they used 8 labels, while we used 12. Considering this, the results of learning on CORE, mapped to the GINCORE schema, and testing on SL-GINCORE, which reached 0.60 micro F1 with the base-sized XLM-RoBERTa model, are promising, showing that mapping to the GINCORE schema gives comparable results to using the CORE schema.

To obtain a deeper insight into the comparability of the GINCO and CORE corpora, we can compare how the F1 scores per labels change when we test the model on another corpus versus when we test it on the same dataset. Figure 3 shows a comparison between the F1 scores per labels for in-dataset experiments with SL-GINCORE and cross-dataset experiments from SL-GINCORE to EN-GINCORE, performed with XLM-RoBERTa. An analysis of these experiments, performed with CroSloEngual BERT, confirmed that differences between label scores occur when learning with any of the two models, and do not depend on the model. The same differences in label scores were also observed in experiments where MT-GINCORE is used instead of SL-GINCORE, which indicates that the language of the dataset does not seem to have a large impact on the results per labels.

As shown in Figure 3, the F1 scores for *News* and *Opinion/Argumentation* are almost the same in both setups, which shows that in regard to these genres, the datasets are comparable enough for the model to generalize from one dataset to the other. The F1 scores are significantly lower in cross-lingual experiments in case of *Promotion*, *Information/Explanation*, *Forum* and *Instruction*. For the labels that are under-represented in the SL-GINCORE, i.e., labels that are on the right side of *Review* in the Figure, it is not possible to ascertain whether the differences between the scores are an indicator that the datasets are not comparable in regard to these labels or that the differences occurred due to chance.

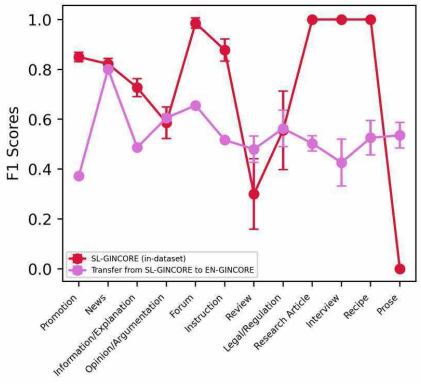


Figure 3: Comparison of average F1 scores per labels between in-dataset experiments and cross-dataset experiments with XLM-RoBERTa. The models were trained on SL-GINCORE, and tested on a) SL-GINCORE (in-dataset experiments) and b) EN-GINCORE (cross-dataset experiments). Labels are ordered according to their frequency in the smallest of the datasets, SL-GINCORE.

As in the in-dataset experiments, experiments with the two Transformer models show that while one dataset combination seems to achieve the best results with one model,

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

it performs differently with the other model. These results once again show the importance of using multiple models on multiple datasets in the experiments to see whether conclusions obtained from experiments with one model are still supported when using another, yet similar model, and how the performance of the models depends on the datasets. While results in terms of micro F1, achieved with XLM-RoBERTa, point to a conclusion that transfer from SL-GINCORE to EN-GINCORE achieves better results than the other direction, macro F1 scores, achieved with XLM-RoBERTa, and both F1 scores, achieved with CroSloEngualBERT, show transfer direction from English to Slovene to be better. However, although the EN-GINCORE dataset is 40 times larger than SL-GINCORE, the transfer from EN-GINCORE to SL-GINCORE does not achieve significantly higher results than the transfer in the other direction when the Slovene dataset is used.

In addition to this, the results show that machine-translating the dataset into English can in some cases improve the results of cross-lingual experiments. In cases where the model was trained on the GINCO datasets, i.e., SL-GINCORE or MT-GINCORE, and tested on the EN-GINCORE dataset, the setup with the machine-translated text achieved slightly lower results than the setup with the original Slovene dataset, SL-GINCORE, in case of XML-RoBERTa, and slightly better results in case of CroSloEngual BERT. However, when the transfer was applied in the other direction, that is, from EN-GINCORE to SL- or MT-GINCORE, machine translating the test instances from Slovene into English resulted in improvements of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores, obtained with CroSloEngual BERT.

5. Conclusions

Following Repo et al. (2021) who showed that good levels of cross-lingual transfer can be achieved by training Transformer models on a large English genre dataset and applying them to datasets in other languages, the goal of this study was to explore whether it is possible to achieve similar results on the Slovene genre dataset. The results revealed to be promising, as despite using a smaller Transformer model and a different schema with more labels than previous work, the results are rather comparable, showing that the English CORE and Slovene GINCO datasets are comparable enough to allow cross-dataset experiments. The XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 in terms of micro and macro F1 scores respectively, are a promising indicator that cross-lingual transfer could be possible in the automatic genre identification task for Slovene as well. Furthermore, high F1 scores achieved with XLM-RoBERTa in monolingual experiments show that automatic genre identification is feasible already with a very small dataset, and that using the GINCORE schema on all datasets gives good results. Moreover, despite the fact that the CORE dataset is 40 times larger than the GINCO dataset, it did not provide consistently significantly better results than the GINCO dataset in either of the setups. We plan to analyze this further by exploring what results can be achieved when smaller portions of CORE are used for training, and by extending the GINCO dataset to analyze whether this further improves the results.

As recently developed trilingual Croatian-Slovene-English CroSloEngual model was shown to be comparable to massively multilingual XLM-RoBERTa model in numerous NLP tasks (see Ulčar et al. (2021)), both models were used in the experiments to analyze their performance in the AGI tasks. The results of both monolingual and cross-lingual experiments showed that despite achieving high results in other common NLP tasks, CroSloEngual BERT seems to be less suitable than XLM-RoBERTa for automatic genre identification.

To improve monolingual and cross-lingual results, we also experimented with translating the Slovene GINCO dataset into English, which is the main language on which the Transformer models were pre-trained. In regard to monolingual experiments, there were no consistent results which would confirm that using an English dataset improves classification. However, when the models were trained on the English EN-GINCORE and tested on MT-GINCORE, i.e., a Slovene dataset, machine-translated into English, this led to improvement of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores for CroSloEngual BERT. This means that machine translating the dataset into the language of another dataset might be beneficial in cross-lingual cross-dataset experiments.

Although monolingual and cross-lingual experiments showed good results also when the models were trained on SL-GINCORE and MT-GINCORE, consisting of less than 1,000 instances, comparisons of F1 scores, reported for each label in different runs and setups, showed that some labels are represented by too few instances to provide reliable results. In the future, we plan to extend the GINCO dataset to assure more reliable results and to further improve the classifiers' performance.

In addition to this, recent work by Rönnqvist et al. (2021) showed that multilingual modeling, where the model was trained on CORE datasets in various languages, resulted in significant gains over cross-lingual modeling, where the model was trained solely on the English CORE dataset. As our research revealed that the CORE and GINCO labels can be successfully mapped to a joint schema, in the future, we plan to extend the experiments to multilingual modeling by training the model on a combination of all CORE datasets and the Slovene GINCO dataset.

Acknowledgments

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

6. References

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2016. Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. Slovene web corpus MaCoCu-sl 1.0. Slovenian language resource repository CLARIN.SI.

Douglas Biber and Jesse Egbert. 2018. *Register variation online*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

DeepL. n.d. Why DeepL? `https://www.deepl.com/en/whydeepl`.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Tomaž Erjavec and Nikola Ljubešić. 2014. The slWaC 2.0 corpus of the Slovene web. *T. Erjavec, J. Žganec Gros (ur.) Jezikovne tehnologije: zbornik*, 17:50–55.

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: *Proceedings of the fifth Web as Corpus workshop*, pages 27–35.

GloWbe. n.d. Corpus of Global Web-Based English (GloWbE): Texts. `https://www.english-corpora.org/glowbe/`.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Taja Kuzman, Mojca Brglez, Peter Rupnik, and Nikola Ljubešić. 2021. Slovene web genre identification corpus GINCO 1.0. Slovenian language resource repository CLARIN.SI.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In: *Proceedings of the Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi, and Sampo Pyysalo. 2020. From web crawl to clean register-annotated corpora. In: *Proceedings of the 12th Web as Corpus Workshop*, pages 14–22.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university Faculty of informatics, Brno, Czech Republic.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In: *16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021*, pages 183–191. Association for Computational Linguistics (ACL).

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2022. American-British-variety Classifier. `https://github.com/macocu/American-British-variety-classifier`.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Matej Ulčar and Marko Robnik-Šikonja. 2020. CroSloEngual BERT 1.1. Slovenian language resource repository CLARIN.SI.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. *arXiv:2107.10614*.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vedrana Vidulin, Mitja Luštrek, and Matjaž Gams. 2007. Using genres to improve search engines. In: *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 45–51.

Ahmad Yulianto and Rina Supriatnaningsih. 2021. Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy*, 9(2):109–127.

3.2.4 Related Paper: Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models

machine learning & knowledge extraction

MDPI

Article

Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models

Taja Kuzman [1,2], Igor Mozetič [1] and Nikola Ljubešić [1,3,*]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia; taja.kuzman@ijs.si (T.K.); igor.mozetic@ijs.si (I.M.)
[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
[3] Center za Jezikovne Vire in Tehnologije Univerze v Ljubljani, 1000 Ljubljana, Slovenia
* Correspondence: nikola.ljubesic@ijs.si

Abstract: Massive text collections are the backbone of large language models, the main ingredient of the current significant progress in artificial intelligence. However, as these collections are mostly collected using automatic methods, researchers have few insights into what types of texts they consist of. Automatic genre identification is a text classification task that enriches texts with genre labels, such as promotional and legal, providing meaningful insights into the composition of these large text collections. In this paper, we evaluate machine learning approaches for the genre identification task based on their generalizability across different datasets to assess which model is the most suitable for the downstream task of enriching large web corpora with genre information. We train and test multiple fine-tuned BERT-like Transformer-based models and show that merging different genre-annotated datasets yields superior results. Moreover, we explore the zero-shot capabilities of large GPT Transformer models in this task and discuss the advantages and disadvantages of the zero-shot approach. We also publish the best-performing fine-tuned model that enables automatic genre annotation in multiple languages. In addition, to promote further research in this area, we plan to share, upon request, a new benchmark for automatic genre annotation, ensuring the non-exposure of the latest large language models.

Keywords: machine learning; text classification; large language models; fine-tuning; automatic genre identification; text genre; web genre

check for updates

Citation: Kuzman, T.; Mozetič, I.; Ljubešić, N. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1149–1175. https://doi.org/10.3390/make5030059

Academic Editors: Jaroslaw Krzywanski, Yunfei Gao, Marcin Sosnowski, Karolina Grabowska, Dorian Skrobek, Ghulam Moeen Uddin, Anna Kulakowska, Anna Zylka and Bachil El Fil

Received: 7 August 2023
Revised: 6 September 2023
Accepted: 7 September 2023
Published: 12 September 2023

1. Introduction

The advent of the World Wide Web provided us with massive amounts of text, useful for information retrieval and the creation of web corpora, which are the basis of many language technologies, including large language models and machine translation systems. To be able to access relevant documents more efficiently, researchers have aimed to integrate genre identification into information retrieval tools [1,2] so that users can specify which genre are they searching for, e.g., a news article, a scientific article, a recipe, and so on. In addition, the web has allowed us easy and fast access to a collection of large monolingual and parallel corpora. Language technologies, such as large language models, are trained on millions of texts. An important factor for achieving a reliable and good performance of these models is assuring that the massive collections of texts are of high quality [3]. The automatic prediction of genres is a robust method for obtaining insights into the constitution of corpora and their differences [4]. This motivates research on automatic genre identification, which is a text classification task that aims to assign genre labels to texts based on their conventional function and form, as well as the author's purpose [5].

The main goal of this paper is to evaluate the out-of-distribution robustness (generalization) of machine learning approaches for text classification in the task of automatic

Mach. Learn. Knowl. Extr. **2023**, *5* 1150

genre identification. Our main focus is on the generalizability of the models across different datasets to assess which model is the most suitable for the downstream task of the automatic annotation of large web corpora with genre information. In summary, this paper presents the following contributions:

- To improve the generalization abilities of classifiers, we create a new genre dataset by merging three manually annotated genre datasets based on a new joint schema. We show that training models on multiple diverse datasets can improve their performance.
- We compare fine-tuned BERT-like Transformer-based models to baseline models that were commonly used for this task in previous research, such as support vector machines (SVMs) and the linear fastText [6] model. We show that Transformer-based language models are state-of-the-art in this task and that earlier machine learning models have poor capabilities in terms of generalization to new datasets.
- We investigate a promising new approach: classifying texts using recent large instruction-tuned GPT Transformer-based language models in a zero-shot setting. As the results reveal it to be a promising approach, we deliberate on the benefits and drawbacks of using BERT-like and GPT-like large language models for text classification tasks.
- In addition, we publish a freely available multilingual BERT-like Transformer-based genre classifier that outperforms other models.
- To promote further research, we introduce a publicly available benchmark with a manually annotated English test dataset: the AGILE (Automatic Genre Identification Benchmark), which can be accessed at https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark (accessed on 6 August 2023).

This paper is organized as follows. In Section 2, we introduce the task of automatic genre identification, its main challenges, and the machine learning approaches used in previous works. In Section 3, we present genre-annotated datasets (Section 3.1) and the models trained and tested on these datasets (Section 3.2). We present baseline models—models that are not Transformer-based—BERT-like models, fine-tuned on different genre datasets, as well as recent GPT models, used in a zero-shot fashion. The results are presented in Section 4. In Section 4.1, we present the performance of the models in the in-domain scenario. In Section 4.2, we analyze the performance of the models in the out-of-domain scenario, and in Section 4.3, we add to the comparison of recent large GPT models that are used in a zero-shot setting. Finally, in Section 5, we conclude this paper with a discussion of the main findings and suggestions for future work.

2. Background

In this section, we present a brief background regarding the task of automatic genre identification, encompassing its inherent challenges. We deliver a literature review of the machine learning experiments reported thus far, which provides a clear overview of the advantages and disadvantages of various machine learning methods, as well as an introduction to manually annotated genre datasets used for training and testing the genre classifiers.

2.1. Impact of Automatic Genre Identification

Having information on the genre of a text is useful for a wide range of fields, including information retrieval, information security, natural language processing, and general, computational and corpus linguistics. While some of the fields mainly base their research on texts that are already annotated with genres, two fields place a greater emphasis on the development of models for automatic genre identification: information retrieval and computational linguistics. With the advent of the World Wide Web, unprecedented quantities of texts became available for querying and collecting. Due to the high cost and time constraints associated with manual annotation, researchers turned their attention toward developing models for automatic genre identification. This approach enables the effortless enrichment of thousands of texts with genre information. The majority of previous works [1,2,7–12] focused on developing models from an information retrieval standpoint.

Mach. Learn. Knowl. Extr. **2023**, *5* 1151

Their objective was to integrate genre classifiers into information retrieval tools, using genre as an additional search query criterion to enhance the relevance of search results [13].

Automatic genre identification has also been researched in the field of computational linguistics, specifically in connection with corpora creation, curation, and analysis. Collecting texts from the web is a rapid and efficient method for gathering extensive text datasets for any language that is present on the web [14]. However, due to the automated nature of this collection process, the composition of web text collections remains unknown [15]. Thus, several previous studies [16–19] researched automatic genre identification with the goal of enriching web corpora with genre metadata. While information retrieval studies mainly focused on a smaller specific set of categories, deemed to be relevant to the users of information retrieval tools, computational linguistics studies focused on developing sets of genre categories that would be able to cover the diversity of genres found on the web. Our paper continues with this line of research, with the aim of providing a genre classifier that can be applied to any text to enrich large web corpora with genre information. This information could provide insights into the textual diversity within the corpora and facilitate genre-based filtering of the collected data. This results in improving the usability of web corpora for corpora linguistics studies, as well as for natural language processing (NLP) applications. Researchers can leverage genre information to select suitable texts for training large language models or employ genre-aware methods for improving NLP tasks, as was done for part-of-speech tagging [20], zero-shot dependency parsing [21], machine translation [22], and the assessment of credibility in web documents [23].

2.2. Challenges in Automatic Genre Identification

To be able to use an automatic genre classifier for the end uses described in the previous subsection, it is crucial that the classifier is robust, that is, it is able to generalize to new datasets. While numerous studies have focused on developing automatic genre classifiers, they were "self-contained, and corpus-dependent" [24]. Most studies reported the results of automatic genre identification based solely on their own datasets, annotated with a specific genre schema. This hinders any comparison between the performance of classifiers from different studies, either in in-dataset or cross-dataset scenarios. In 2010, a review study encompassed all the main genre datasets developed up to that time [1,2,7,16,25,26]. It showed that if we train a classifier on the training split and evaluate it on the test split from the same genre dataset, the results show a rather good performance of the model. However, cross-dataset comparisons, that is, testing the classifiers on a different dataset, revealed that the classifiers are incapable of generalizing to a novel dataset [27]. The applicability of these models for end use is thus questionable.

To address concerns regarding classifier reliability and generalizability, in the past decade, researchers have invested considerable effort in refining genre schemata, genre annotation processes, and dataset collection methods [17,19,28–30]. These studies addressed the difficulties with this task, which impact both manual and automatic genre identification. The main challenges identified were (1) varying levels of genre prototypicality in web texts, (2) the presence of features of multiple genres in one text, and (3) the existence of texts that might not have any discernible purpose or features [1,31].

Recently, three approaches have proposed genre schemata specifically designed to address the diversity of web corpora: the schemata of the English CORE dataset [17], the Slovenian GINCO dataset [19], and the English and Russian Functional Text Dimensions (FTD) datasets [28]. All of them use categories that cover the functions of texts, and some of the categories have similar names and descriptions, which suggests that they might be comparable. This question was partially addressed by Kuzman et al. [32] who explored the comparability of the CORE and GINCO datasets by mapping the categories to a joint schema and performing cross-dataset experiments. Despite the datasets being in different languages, the results showed that they are comparable enough to allow cross-dataset and cross-lingual transfer. Similarly, Repo et al. [33] reported promising cross-lingual and cross-dataset transfer when using the CORE dataset and Swedish, French, and Finnish

Mach. Learn. Knowl. Extr. **2023**, *5* 1152

datasets annotated with the CORE schema. Training a classifier on multiple datasets not only improves its cross-lingual capabilities but also assures better generalizability to a new dataset by mitigating topical biases [34]. This is important since, in contrast to topic detection, genre classification should not rely solely on lexical information such as keywords. The classification of genre categories necessitates the identification of higher-level patterns embedded within the texts, which often stem from textual or syntactic characteristics that are not directly linked to the specific topic addressed in the document.

2.3. Machine Learning Methods for Automatic Genre Identification

The machine learning results reported in the existing literature are dependent on a specific dataset that the researchers used for training and testing the classifier, a machine learning technology of their choosing, and are reported using different metrics. Thus, it remains unclear which machine learning method is the most suitable for automatic genre identification, especially in regard to its generalizability to novel datasets.

In previous research, the choice of machine learning model was primarily determined by the progress achieved in developing machine learning technologies up to that particular point in time. Before the emergence of neural networks, the most frequently used machine learning method for automatic genre identification was support vector machines (SVMs) [27,35–37], which continues to be valuable for analyzing which textual features are the most informative in this task [38,39]. Other non-neural methods, including discriminant analysis [40,41], decision tree classifiers [8,42], and the Naive Bayes algorithm [10,40], were also used for genre classification. Multiple studies searched for the most informative features in this task. They experimented with lexical features (words, word or character n-grams), grammatical features (part-of-speech tags) [31,38], text statistics [8], visual features of HTML web pages such as HTML tags and images [43–45], and URLs of web documents [10,46,47]. However, the results for the discriminative features varied across studies and datasets. One noteworthy limitation of non-neural models lies in their reliance on feature selection, which necessitates a new exploration of suitable features for every genre dataset and machine learning method. Furthermore, as the choice of features relies heavily on the dataset, this hinders the model's ability to generalize to new datasets or languages [48].

Then, the developments in the NLP field shifted the focus to neural networks, which showed very promising performance in this task. One of the main advantages of these models is that their architecture involves a machine-learned embedding model that maps a text to a feature vector [48]. Thus, manual feature selection was no longer needed. Traditional methods were outperformed in this task by the linear fastText [6] model [49]. However, its performance diminishes when confronted with a small dataset encompassing a larger set of categories [19].

This is where deep neural Transformer-based BERT-like language models proved to be extremely capable, surpassing the fastText model by approximately 30 points in micro- and macro-F1 [19]. Transformer is a neural network architecture, based on self-attention mechanisms, which significantly improve the efficiency of training language models on massive text data [50]. Following the introduction of this groundbreaking architecture, numerous large-scale Transformer-based Pre-Trained Language Models (PLMs) arose. PLMs can be divided into autoregressive models, such as GPT (Generative Pre-Trained Transformer) [51] models, and autoencoder models, such as BERT (Bidirectional Encoder Representations from Transformers) [52] models [48]. The main difference between them is the method used for learning a textual representation: while autoregressive models predict a text sequence word by word based on the previous prediction, autoencoder models are trained by randomly masking some parts of the text sequence or corrupting the text sequence by replacing some of its parts [48]. While autoregressive models have been mainly used for generative tasks, autoencoder models have demonstrated remarkable capabilities when fine-tuned to categorization tasks, including automatic genre identification. Thus, some recent studies have used BERT-like Transformer-based language models, which were

Mach. Learn. Knowl. Extr. **2023**, *5* 1153

pre-trained on massive amounts of text collections and fine-tuned on genre datasets. These models were shown to be capable of achieving good results even when trained on only around a thousand texts [19] and provided with only the first part of the documents [53]. Models trained on approximately 40,000 instances and models trained on only a few thousand instances have demonstrated comparable performance [32,33]. These results indicate that massive amounts of data are no longer essential for the models to acquire the ability to differentiate between genres. Additionally, fine-tuned BERT-like models have exhibited promising performance in cross-lingual and cross-dataset experiments [32,33,54].

Among the available monolingual and multilingual autoencoder models, the multilingual XLM-RoBERTa model [55] has proven to be the most appropriate for the task of automatic genre identification. It has outperformed other multilingual models and achieved comparable or even superior results to monolingual models [32,33,54]. Nevertheless, despite the superior performance exhibited by fine-tuned BERT-like Transformer models, a considerable proportion of instances—up to a quarter—continue to be misclassified. The most recent in-dataset evaluations of fine-tuned BERT-like models on the CORE [17] and GINCO [19] datasets yielded micro-F1 scores ranging from 0.68 to 0.76 [19,33,53]. This demonstrates that this text categorization task is much more complex than tasks that mainly depend on lexical features such as topic detection, where the state-of-the-art BERT-like models achieve an accuracy of up to 0.99 [48].

While BERT-like models demonstrate exceptional performance in this task, they still require fine-tuning using a minimum of a thousand manually annotated texts. The process of constructing genre datasets presents several challenges, which involve defining the concept of genre, establishing a genre schema, and collecting instances to be annotated. Additionally, it is crucial to provide extensive training to annotators to ensure a high level of inter-annotator agreement. Manual annotation is a resource-intensive endeavor, demanding substantial time, effort, and financial investment. Furthermore, despite great efforts to assure reliable annotation, inter-annotator agreement in annotation campaigns often remains low, consequently impacting the reliability of the annotated data [1,17,30].

Recent advancements in the field have shown that using instruction-tuned GPT-like Transformer models, more specifically, GPT-3.5 and GPT-4 models [56], prompted in a zero-shot or a few-shot setting, could make these large manual annotation campaigns redundant, and only a few hundred annotated instances would be needed for testing the models. These recent GPT models have been optimized for dialogue based on reinforcement learning with human feedback [57]. While they were primarily designed as a dialogue system, there has recently been a growing interest among researchers in investigating their capabilities in various NLP tasks such as sentiment analysis, textual similarity, natural inference, named-entity recognition, and machine translation. While some studies have shown that the GPT-3.5 model was outperformed by the fine-tuned BERT-like large language models [58], it exhibited state-of-the-art results in stance detection [59], high performance in implicit hate speech categorization [60], and competitive performance in machine translation of high-resource languages [61]. Building upon these findings, a recent pilot study [62] explored its performance in automatic genre identification. The study used the model through the ChatGPT interactive interface, as at the time of the research, the model was not yet available through an API. Used in a zero-shot setting, the performance of the GPT-3.5 model was compared to that of the XLM-RoBERTa model [55], fine-tuned on genre datasets. Remarkably, the GPT-3.5 model outperformed the fine-tuned genre classifier and exhibited consistent performance, even when applied to Slovenian, an under-resourced language. Furthermore, OpenAI has recently introduced the GPT-4 model, which was shown to outperform the GPT-3.5 model family and other state-of-the-art models across a range of NLP tasks [63]. These findings suggest the significant potential of using GPT-like language models for automatic genre identification.

Mach. Learn. Knowl. Extr. **2023**, *5* 1154

3. Materials and Methods

In this section, we introduce the manually annotated genre datasets, which are presented in Section 3.1. They are used for training and testing the genre classifiers, which are presented in Section 3.2.

3.1. Genre Datasets

In the experiments, we use five manually annotated datasets in two languages—English and Slovenian.

In Section 3.2.2, we experiment with training and testing pre-trained BERT-like models on three previously existing genre datasets, each using their own genre schema:

- The English Functional Text Dimensions (FTD) dataset [28];
- The Corpus of Online Registers of English (CORE) dataset [17];
- The Slovenian Genre Identification Corpus (GINCO) dataset [19].

We chose these datasets as they aim to represent the entire diversity of genres found on the web, which is aligned with our goal. While there exist multiple other genre-annotated datasets, they were not included in our experiments due to one or more of the following reasons: (1) they were collected with a different aim and consist of a small set of specific genres that do not cover the diversity of genres found on the web (e.g., the Genre-KI-04 [1] and the 7-Web Genre Collection [26] datasets); (2) they are not available (e.g., 20-Genre Collection (MGC) [2], Hierarchical Genre Collection (HGC) [7], SANTINIS-ML [64] and Leeds Web Genre Corpus [29]); and (3) they are not in the English or Slovenian languages (e.g., the French FreCORE and Swedish SweCORE datasets [65], the Finnish FinCORE dataset [66], and the Russian LiveJournal [34] and FTD datasets [28]). In this paper, we limit our multilingual experiments to two languages to obtain controlled insights into the models' multilingual capabilities. In future work, we plan to extend our experiments to additional languages.

In addition, in this paper, we introduce two newly created datasets:

- The English and Slovenian X-GENRE dataset, which consists of samples from the FTD, CORE, and GINCO datasets, and introduces the X-GENRE schema that merges the schemata from previous datasets into one single schema;
- The English EN-GINCO test set, which was manually annotated with the X-GENRE schema.

Lepekhin and Sharoff (2022) [34] demonstrated that combining datasets improves performance, and previous experiments have demonstrated some level of compatibility between the GINCO and CORE datasets [32]. Following these findings, we combine the FTD, GINCO, and CORE datasets into a new multilingual genre dataset to improve the generalizability of the trained models. The X-GENRE dataset thus consists of three datasets in two languages collected from different sources, in different time periods, and using different collection methods. To analyze the in-domain performance of various machine learning methods, we train and test the classifiers on this dataset. In addition, we introduce a new test set (EN-GINCO) to analyze the out-of-domain robustness of the classifiers. The X-GENRE and EN-GINCO datasets are also suitable for the evaluation of the performance of the recent GPT models, as the annotations have never been published on the web. This ensures that the results cannot be impacted by a data leakage between the GPT-3.5 and GPT-4 training and test sets.

Table 1 presents an overview of the genre datasets that were used in our experiments. The datasets are described in more detail in the following subsections and in Appendix A.

Mach. Learn. Knowl. Extr. **2023**, *5* 1155

Table 1. A comparison of genre datasets, on which we fine-tune and evaluate XLM-RoBERTa models, in terms of language, number of genre labels, and number of texts (further divided into training, evaluation, and test splits).

Dataset (Lang)	Labels	Texts (Train-Dev-Test)
CORE-main (EN)	9	17,094 (10,256-3419-3419)
CORE-sub (EN)	37	15,895 (9537-3179-3179)
GINCO-reduced (SL)	17	965 (579-193-193)
GINCO-downcast (SL)	9	1002 (601-201-200)
FTD (EN)	10	1415 (849-283-283)
X-GENRE (SL + EN)	9	2956 (1772-592-592)
EN-GINCO (EN)	9	272 (test only)

3.1.1. Genre Datasets from Previous Works

The **CORE** dataset [17] consists of 48,420 English web texts. It is annotated with a two-level hierarchical schema, consisting of 8 main categories and 47 more specific subcategories (see Appendix A for more information on the text collection and annotation procedure). To use the dataset in the single-label classification task, we conduct separate experiments using each annotation level independently. Moreover, our preliminary experiments and the findings of Laippala et al. (2022) [53] show that the model's performance reaches a plateau and exhibits no significant improvements beyond the utilization of 30% of the available CORE instances. Following these findings, we opt to use a subset consisting of 40% of the CORE dataset. When extracting the subset, we perform a stratified split to ensure the preservation of the original label distribution. Thus, the final versions of CORE that we use in the experiments are as follows:

- CORE-main: a sample of the CORE dataset, consisting of 17,094 texts. For labels, we use the 9 CORE main categories.
- CORE-sub: a sample of the CORE dataset, comprising 15,895 texts, annotated with CORE subcategories. We only use labels that are represented by more than 10 instances, resulting in the final label set of 37 categories.

The **GINCO** dataset [19] consists of 1002 Slovenian web texts. The dataset was manually annotated with 24 genre categories from the GINCO schema (see Appendix A for more information on the text collection and annotation procedure). The GINCO schema is based on the subcategory level of the CORE schema but has a lower granularity of categories to facilitate manual annotation and improve the performance of the genre classifiers. Based on the findings reported in a previous study [19], which highlight that the performance of the classifiers could be further enhanced by applying some modifications to the GINCO schema, we experiment with two versions of the GINCO dataset:

- GINCO-reduced: only labels, represented with more than 10 instances, are used. This intervention amounts to 37 discarded texts. The final dataset has 17 labels and 965 texts.
- GINCO-downcast: the original GINCO labels are merged into 9 broader categories, and no texts are discarded.

The English **FTD** dataset [28] consists of 1562 texts. The annotation employed a multi-label approach, where the presence of labels was evaluated on a scale ranging from 0 to 2 (see Appendix A for more details on the text collection and annotation procedure). The dataset, which we use for the experiments, is annotated with 10 genre categories and an "unsuitable" category. To use the multi-label schema for a single-label classification, texts annotated with a 2 (from a scale from 0 to 2) are considered as belonging to that particular category. For the experiments, we discard texts belonging to multiple labels and texts annotated as "unsuitable". Thus, the final dataset used in our experiments consists of 1415 instances annotated with 10 labels.

3.1.2. Newly Introduced Genre Datasets

In line with previous works, which demonstrated improved performance when a model was trained on multiple datasets [34], even in multiple languages [33,54], we create a novel genre dataset by merging three commonly used datasets. The **X-GENRE** dataset consists of 2956 instances, out of which:

- 893 texts are from the Slovenian GINCO [19] dataset;
- 1050 texts are from the English FTD [28] dataset;
- 1013 texts are sampled from the English CORE [17] dataset.

The three genre datasets use distinct genre schemata with different label names and label set granularity. Thus, to merge them, it is necessary to develop a joint schema. Certain label names may suggest that they could be directly mapped, e.g., *A7* (*instruction*) (FTD), *Instruction* (GINCO), and *How-To* (CORE subcategory). However, the texts associated with these labels do not necessarily possess comparable genre characteristics. The labels from different datasets could be defined differently in the annotation guidelines, and there might be additional variations in how the annotators interpreted them. This is why we opt for a data-based approach: we train separate genre classifiers on the FTD dataset, GINCO-downcast dataset, and CORE-sub dataset, as described in Section 3.2.2. Then, we apply each classifier to the other two datasets and analyze, for each dataset, how frequently a true label from one schema matches a label predicted by a classifier that was trained on the other dataset and its schema. For instance, we apply the FTD classifier—a classifier, fine-tuned on the FTD dataset—on the GINCO dataset, and analyze the frequency of matches between the GINCO true labels and FTD predicted labels. Some labels of each of the datasets were found to be dispersed across multiple categories in the other dataset and could not be mapped to a single label. We thus discard these labels and do not use texts annotated with them in the final X-GENRE dataset. The final joint schema, named the X-GENRE schema, comprises 9 labels, which cover 22 of the original 24 GINCO categories, 8 of the 10 original FTD categories, and 22 out of the 47 CORE subcategories. The final X-GENRE labels are *Information/Explanation*, *Instruction*, *Legal*, *News*, *Opinion/Argumentation*, *Promotion*, *Prose/Lyrical*, *Forum*, and *Other* (for more details, see the definitions of the labels in Table A1 in Appendix A). The mapping of the schemata to the joint schema is shown in Figure 1.

The merged dataset has multiple advantages:

- As it consists of multiple datasets, it provides a greater variety of instances, consequently enhancing the generalization capabilities of models trained on the dataset.
- It consists of two languages and allows the analysis of the multilingual performance of genre classifiers.
- Its English component outweighs the Slovenian component, representing two-thirds of the instances. This is an additional advantage of the dataset, as it allows us to examine the performance of models when confronted with a high-coverage language (English) in contrast to an under-resourced language—a language that is less represented in the data (Slovenian).

In supervised machine learning, it is customary to partition a dataset into three distinct subsets: training, evaluation, and testing. The training split is used to train the model, whereas the test split is employed to evaluate its performance. It is crucial that these splits are non-overlapping, meaning that no instance appears in more than one split. This ensures that the model's ability to generalize to new instances is assessed, rather than its capacity to simply memorize the instances in the training split. We assess the models in both in-domain and out-of-domain scenarios. In the case of the in-domain experiments, the models are trained and tested on splits derived from the same dataset. In contrast, in the out-of-domain experiments, the models are trained on one dataset and tested on another. The evaluation split is used to evaluate the model's performance when searching for the optimal hyperparameters for training. All the models, compared in Section 4, are trained on the training subset and tested on the test subset of the X-GENRE dataset (for the in-domain experiments). The X-GENRE dataset is split into training, evaluation, and testing sets in a

Mach. Learn. Knowl. Extr. **2023**, *5* 1157

60:20:20 manner (1772:592:592 texts). Stratified splitting is employed to ensure a consistent label distribution across all subsets. In addition, we maintain an equal representation of instances from the FTD, CORE, and GENRE datasets within each split to preserve a consistent distribution of English and Slovenian instances throughout the subsets.

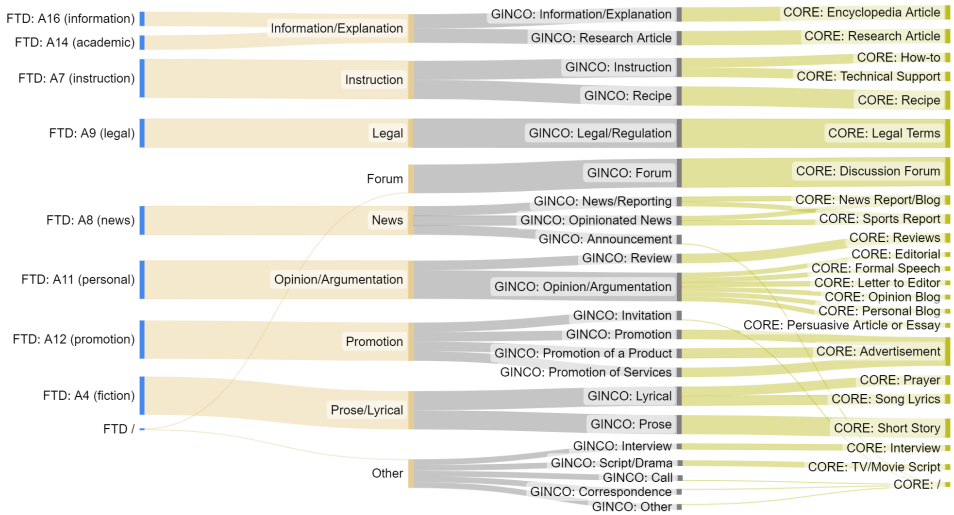


Figure 1. A diagram showing how labels from the FTD [28] (first column), GINCO [19] (third column), and CORE [17] (fourth column) schemata are mapped to the new joint X-GENRE schema (second column).

The **EN-GINCO** test set is prepared with the goal of testing the out-of-domain performance of the models. The test set consists of 272 English texts. They are sampled from the English web corpus enTenTen20 [67] and annotated by two expert annotators. The EN-GINCO dataset uses the same schema as the X-GENRE dataset, which enables testing classifiers, trained on the X-GENRE dataset, in an out-of-distribution scenario (the differences between the label distributions in the datasets are shown in Table A2 in Appendix A).

Although the texts from the X-GENRE and EN-GINCO datasets originate from the web, the annotations have not been published, which makes the datasets suitable for the evaluation of GPT-3.5 and similar models, as it is certain that these labeled datasets were not included in the model's training data.

3.2. Models

In this subsection, we present the machine learning architectures that we compare in the in-domain and out-of-domain experiments:

- Dummy classifier;
- Naive Bayes classifier;
- Support Vector Machine (SVM);
- fastText shallow neural model;
- Multilingual Transformer-based base-sized XLM-RoBERTa model;
- Autoregressive instruction-tuned GPT-3.5 and GPT-4 models.

In Section 3.2.1, we present the machine learning technologies that precede the Transformer models. These technologies serve as our baseline models for comparison. In Section 3.2.2, we conduct an evaluation of the XLM-RoBERTa models, fine-tuned on various genre datasets, presented in Section 3.1. Our objective is to ascertain the genre dataset that yields the optimal performance in this task. The best-performing model is selected for the in-domain and out-of-domain experiments, as described in Section 4. Finally, in Section 3.2.3, we present the GPT-3.5 and GPT-4 models, which are instruction-based and do not necessitate fine-tuning on a genre dataset.

Mach. Learn. Knowl. Extr. **2023**, *5* 1158

3.2.1. Baseline Models

A significant portion of the existing literature on automatic genre identification predominantly relies on non-Transformer-based machine learning methods, primarily due to the unavailability of Transformer models during the time of research. Consequently, we incorporate these non-Transformer models into our comparative analysis, allowing us to gain insights into their efficacy in relation to the state-of-the-art Transformer models in various scenarios, including in-domain, out-of-domain, and multilingual contexts. We use the following models, provided through the Scikit-Learn library [68]:

- **Dummy Classifier**: The model serves as a simple baseline to reveal what the model's performance would be without being trained on the labeled texts. The classifier uses the stratified strategy, which means that the predictions are based on the information on the label distribution in the training data.
- **Naive Bayes Classifier**: This probabilistic machine learning algorithm learns the statistical relationships between the words present in the documents, also taking into account their frequency and the corresponding genre categories. We use the complement Naive Bayes implementation, which is particularly suited to imbalanced multi-class datasets [69].
- **Logistic Regression Classifier**: This algorithm models the relationship between the features (words) and the probability of belonging to a category using a logistic function. The cross-entropy loss is used to determine the most probable class. We use the implementation based on the limited-memory BFGS (L-BFGS) solver [70], which is suitable for addressing the complexities of multi-class classification.
- **Support Vector Machine (SVM)**: The SVM model is a linear classifier that determines the boundaries between classes in the form of a separating hyperplane. Its efficacy is particularly notable in high-dimensional spaces, making it highly applicable in the context of text categorization tasks, where the feature set can encompass the entire dataset vocabulary. SVMs have successfully been applied in multiple automatic genre identification studies [27,36,38,39], achieving a micro-F1 of up to 0.75 on the subset of the CORE dataset [38]. In this study, we employ the SVC implementation with the linear kernel, which supports multi-class categorization.

In addition to the non-neural models, we experiment with the **fastText model** [6], which is a shallow neural network. The model has one hidden layer, where the word embeddings are created and averaged into a text representation. The document representation is then fed into a linear classifier, which predicts the genre labels.

As is common in text categorization tasks, the inputs to the models are documents (text strings), along with their target labels. Neural models do not require any feature selection but represent the entirety of the textual information. This is why we also do not perform feature selection in the case of non-neural models, and we use the lexical text representation—words in running text—as features. The non-neural models require a numeric representation of the documents as input. Thus, we represent the documents using the TF-IDF (Term Frequency–Inverse Document Frequency) algorithm. This method represents the document as a vector with values for all words in the dataset. It assigns a higher weight to words occurring more frequently in a small number of texts and a lower weight to words that are present in a large number of texts. The algorithm is implemented as a TfidfVectorizer in the Scikit-Learn library [68].

We perform a hyperparameter search for all the models to ensure that the optimal hyperparameters are used (see Appendix B for more details on the hyperparameters). The models are trained on the training split of the X-GENRE dataset.

3.2.2. Fine-Tuned XLM-RoBERTa Models

To evaluate the performance of the fine-tuned autoencoder models in this task, we use the massively multilingual XLM-RoBERTa model [55], which is available on the Hugging Face model repository (https://huggingface.co/xlm-roberta-base (accessed on 30 November 2022)). Prior studies that compared multiple monolingual and multilingual models

Mach. Learn. Knowl. Extr. **2023**, *5* 1159

identified this particular model as the most suitable model for the task of automatic genre identification [32,33,54]. The XLM-RoBERTa model is a Transformer-based large language model that was pre-trained with the masked language modeling objective, where some of the words in the text are masked and the model is trained to predict the correct word. A Transformer is a neural network architecture, which includes self-attention mechanisms that significantly improve its performance on massive text data [50]. The XLM-RoBERTa model was pre-trained on the CommonCrawl multilingual data [52], which comprise 167 billion tokens in 100 languages, out of which Slovenian is represented by 1.7 billion tokens. We use the base-sized model that has 12 hidden layers and 768 hidden states. To use the pre-trained model for automatic genre identification, we fine-tune it on a genre dataset using the Simple Transformer library (available at https://simpletransformers.ai/ (accessed on 30 November 2022)) on an NVIDIA V100 GPU. The input to the model is running text, which is transformed into an initial numerical representation using the model's tokenizer and embedding model. No additional pre-processing of the input text or feature selection is performed.

To investigate which genre dataset provides the best results, we conduct experiments involving fine-tuning the XLM-RoBERTa model on various genre datasets, namely the FTD [28] dataset, the CORE [17] dataset, the GINCO [19] dataset, and our newly prepared X-GENRE dataset. We use two versions of the GINCO dataset and two versions of the CORE dataset. GINCO-downcast and GINCO-reduced share the same instances but are labeled with different simplifications of the original GINCO schema. Likewise, CORE-sub and CORE-main are derived from the CORE dataset, with labeling based on either the main CORE labels or the CORE subcategories. For more details on the datasets, refer to Section 3.1.

The datasets are split into training, evaluation, and testing subsets in a 60:20:20 manner, and the subsets are stratified based on the label distribution. We perform a hyperparameter search for each model, where we train it on the training split and evaluate it on the evaluation split. More details on the final hyperparameters used are described in Appendix B.

Table 2 compares the XLM-RoBERTa models, fine-tuned on different genre datasets. The models are evaluated on the test split using the micro-F1 and macro-F1 metrics to measure the instance- and class-level performance. The evaluation of the models is conducted in an in-dataset scenario, where each model is tested on the test split that comes from the same dataset as the training split used for fine-tuning. As the genre datasets are labeled with different genre schemata, this means that the models are trained and tested using different category sets. The results show that the best-performing model is the one fine-tuned on the X-GENRE dataset, which consists of instances from the FTD, CORE, and GINCO, merged based on a joint schema. This outcome supports our hypothesis that training genre models on multiple datasets improves the results, which is also consistent with the earlier findings by Repo et al. (2021) [33] and Lepekhin and Sharoff (2022) [34]. The model fine-tuned on the X-GENRE dataset achieves a micro-F1 score of 0.80 and a macro-F1 score of 0.79.

The FTD and GINCO-downcast datasets also demonstrate suitability for automatic genre identification, with micro- and macro-F1 scores above 70. The results based on the GINCO-downcast dataset also reveal a remarkable ability of the multilingual XLM-RoBERTa model to classify Slovenian texts, despite Slovenian representing a very small portion of the XLM-RoBERTa pre-training dataset. The model trained on the English FTD dataset achieves micro- and macro-F1 scores of only one to two points higher than those achieved by the model trained on the Slovenian GINCO dataset, whereas the model trained on the CORE-main dataset achieves high micro-F1 and macro-F1 scores of 0.62, indicating its limitations in identifying less frequent categories. This discrepancy can be attributed to the significant class imbalance present in the CORE-main dataset—although the dataset is labeled with 9 categories, more than 90% of instances fall under the 5 most frequent categories. The difference between the datasets in terms of category distribution is shown in Figure 2.

Mach. Learn. Knowl. Extr. **2023**, *5* 1160

Table 2. In-domain results of various XLM-RoBERTa-based genre classifiers, fine-tuned on different genre datasets and their own schemata. The models are trained on the training split and evaluated on the test split from the same dataset. The results are ordered based on the macro-F1 scores.

Dataset (No. of Labels)	Micro-F1	Macro-F1
X-GENRE (9)	**0.80**	**0.79**
FTD (10)	0.74	0.74
GINCO-downcast (9)	0.73	0.72
CORE-main (9)	0.75	0.62
GINCO-reduced (17)	0.59	0.47
CORE-sub (37)	0.66	0.39

Additionally, the results in Table 2 highlight that a higher granularity of a genre schema has an adverse effect on genre classification. The GINCO-reduced dataset, labeled with 17 categories, achieves 14 points less in the micro-F1 score and 25 points less in the macro-F1 score than the GINCO-downcast dataset, which uses 9 labels. Similarly, the CORE-sub dataset, comprising 37 genre categories, is shown to be unsuitable for training genre classifiers, as the model achieves a mere 0.39 in terms of the macro-F1 score, indicating an inability to differentiate between such a large number of labels, particularly the less frequent ones.

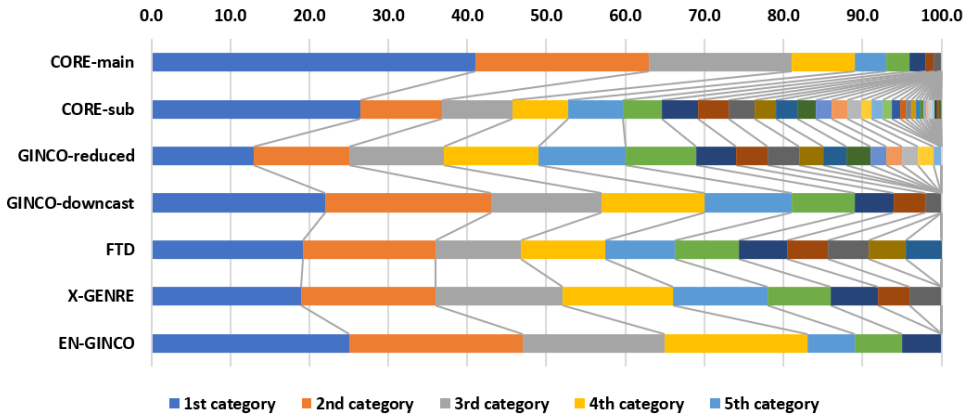


Figure 2. The comparison of category distributions in genre datasets in terms of percentages of instances belonging to each category, from the most frequent category (1st category) to the least frequent one.

As the results show that the X-GENRE dataset is the most suitable for modeling genres, we use this dataset for the comparison between different machine learning models in the in-domain and out-of-domain scenarios presented in Section 4.

3.2.3. Instruction-Tuned GPT Models

To analyze the performance of GPT models in this task, we use the recent GPT-3.5 [56] and GPT-4 [63] models, provided by OpenAI, which have shown very competitive performance in various categorization tasks [59–63]. Generative Pre-Trained Transformer language models (GPTs) are pre-trained to predict the next token in a text. The GPT-3.5 and GPT-4 models are said to be trained on massive multilingual web text collections; however, the details of the datasets are not available. After pre-training, the models were fine-tuned to follow instructions with additional data based on the reinforcement learning with human feedback [57] algorithm. More precisely, the models were first fine-tuned on a dataset of prompts and human-generated answers. After fine-tuning, a new dataset was created, consisting of a sample of prompts and multiple answers provided by the models for each prompt. The annotators then rated the models' answers based on their suitability. The models' performance was then optimized based on a reward model, which

Mach. Learn. Knowl. Extr. **2023**, *5* 1161

was trained to predict which of the outputs was rated the highest by humans. The reward function was optimized with reinforcement learning using the proximal policy optimization algorithm [71].

The GPT-4 model represents a more recent iteration of the large GPT models, incorporating further optimization methods during the post-training alignment process. The primary objective of these improvements is to enhance the quality and accuracy of the model's responses, surpassing those of its predecessor, GPT-3.5, while also ensuring greater stability of its behavior [63]. The model has been shown to outperform the GPT-3.5 model and other large language models in various NLP tasks, including commonsense reasoning [63]. Furthermore, the GPT-4 model is multimodal, meaning that it can process both text and image inputs. The developers of the GPT-3.5 and GPT-4 models have not disclosed any further details regarding the training methods, datasets, or architectures employed.

For the experiments, we use the GPT-3.5 Turbo and GPT-4 models and the chat completion endpoint through the OpenAI API. The models are used in a zero-shot fashion, meaning that we prompt the models as they are and do not fine-tune them on the X-GENRE dataset, in contrast to the other models in the comparison. At the same time, the models do receive some context on the automatic genre identification task via our prompting. The prompt consists of the instruction "*Please classify the following text according to genre (defined by function of the text, author's purpose and form of the text). You can choose from the following classes: News, Legal, Promotion, Opinion, Instruction, Information, Literature, Forum, Other. The text to classify:*", as well as the text from the EN-GINCO or X-GENRE datasets to be classified (an instance of the prompt is provided in Appendix C). The X-GENRE dataset also contains Slovenian instances. In these cases, the instruction remains the same (written in English), while the text to be classified is provided in Slovenian in its original form. The models have a limitation on the number of tokens from the prompt and the completion that can be processed for one request so the texts are truncated to the first 200 words. The prompt uses the X-GENRE labels; however, we shorten some of the label names to one word that consists of only one token to facilitate the parsing of the models' outputs. The details of the hyperparameters used are described in Appendix B.

One of the disadvantages of the GPT-3.5 and GPT-4 models is that as generative models, they are not constrained to output a label from a predefined genre set. This requires careful construction of the prompt to minimize the occurrence of extraneous labels. In addition, we post-process the results to ensure consistency with the original label set. Specifically, we correct the predicted labels that are very similar to the original labels, e.g., "Other(" to "Other", "Instruction/" to "Instruction", and so on. In addition, predictions not belonging to the label list, such as *Interview*, *Condolence*, *Religious*, and *Policy*, are consolidated under the label *Other*. However, it is important to note that the incidence of such cases was relatively low. Across all the experiments conducted on the EN-GINCO and X-GENRE test splits, the GPT-3.5 model generated 13 labels that were not part of the original label list, and they were assigned to a total of 30 texts, constituting a mere 3% of both datasets combined. Conversely, the occurrence of this issue was less frequent with the GPT-4 model, which generated 5 labels outside the label set. These occurrences were observed in 7 cases, which represents 0.8% of instances from both datasets.

4. Results

In this section, we present the results of the in-domain, out-of-domain, and zero-shot experiments. In the in-domain scenario, outlined in Section 4.1, the models are trained on the training split of the X-GINCO dataset and tested on the test split of the same dataset. These experiments aim to evaluate the performance of the models inside the same dataset and reveal important differences between the capacities of the models. The X-GENRE dataset encompasses instances in two languages, namely English and Slovenian. Thus, it enables a comparison of the models' performance between a high-resource language, predominantly represented in the dataset, and a low-resource language.

Mach. Learn. Knowl. Extr. **2023**, *5* 1162

In Section 4.2, we investigate the performance of the models in the out-of-domain scenario. This assessment sheds light on their generalization ability, also known as out-of-distribution performance or robustness. We use the same models as in Section 4.1, that is, the models, trained on the X-GINCO dataset. However, in this case, the models are tested on a novel dataset—the EN-GINCO dataset. In addition, in Section 4.3, we expand our analysis by incorporating a novel approach to text categorization, employing GPT models guided by instructions in the form of prompts. As these models were not specifically trained on a genre dataset, we regard their performance as zero-shot.

Across all three scenarios, we assess the models' performance using the following common evaluation measures for text classification tasks [72]: accuracy, micro-F1 score, and macro-F1 score. One should note that accuracy is less informative in the case of imbalanced datasets, which is a characteristic of most genre datasets (see Section 3.1). However, we include accuracy to enable comparison with previous studies [27,73,74]. The main metrics used for reporting the results are the micro- and macro-F1 scores. The micro-F1 score considers the global recall and precision across categories and thus reports the classifiers' per-instance performance. Consequently, it is more influenced by the most frequent label. This is why the main metric on which we base the interpretation of the results is the macro-F1 score. This measure is computed using the arithmetic mean of the per-class F1 scores. Thus, it shows the classifiers' performance across labels and is not influenced by the label distribution.

For the classifiers to be applicable to new data, it is crucial that they are able to reliably classify all the classes in the label set. Thus, we emphasize the macro-F1 scores, as they provide insights into the models' performance at the class level. McNemar's test [75] is employed to assess the statistical significance of the differences between the two top-performing models in each scenario. Rather than comparing their accuracy rates, this test focuses on whether the models exhibit the same level of disagreement by examining situations where one model is correct but the other is not. This particular statistical test is chosen due to its suitability in evaluating deep learning models for which multiple runs of experiments would be computationally expensive [76].

4.1. In-Domain Performance

Table 3 shows the performance of the baseline models and the fine-tuned XLM-RoBERTa model in the in-dataset scenario. This means that the models trained on the training split of the X-GENRE dataset are tested on the test split from the same dataset. As shown in the table, the fine-tuned XLM-RoBERTa model achieves a micro-F1 score of 0.80 and a macro-F1 score of 0.79, significantly outperforming the baseline classifiers. We assess the statistical significance of the difference between the XLM-RoBERTa model and the second-best performing model, the Logistic Regression classifier, using McNemar's test. The results indicate that the observed difference is statistically significant, with a p-value below 0.0005. Most of the baseline models, that is, the Logistic Regression classifier, the SVM model, and the fastText model achieve micro- and macro-F1 scores of between 0.64 and 0.67. The Naive Bayes classifier is shown to be the least capable of identifying genres. As the fine-tuned XLM-RoBERTa model proved to be the best-performing model in this task, we have made the model publicly available on the HuggingFace repository.

Since the X-GENRE dataset encompasses instances in two languages (English and Slovenian), it offers an opportunity to compare the performance of the models on both a high-resource language, which is prominently represented in the dataset, and a low-resource language. We partition the instances of the X-GENRE test split based on their respective languages and evaluate the models' performance separately for each language. Table 4 shows the differences in the scores achieved on the English part and those obtained on the Slovenian part of the X-GENRE test split. The results reveal that all classifiers exhibit inferior performance on the Slovenian portion of the X-GENRE test split. The Naive Bayes classifier is the least suitable for multilingual datasets, as evidenced by its particularly poor performance in terms of the macro-F1 scores. When applied to Slovenian instances, it

Mach. Learn. Knowl. Extr. **2023**, *5* 1163

achieves a macro-F1 score of 0.18, which is 36 points lower than its performance on English instances. The difference between the models is also significant based on McNemar's test in the case of all other baseline classifiers, with macro-F1 scores ranging from 25 to 29 points lower. In contrast, the fine-tuned XLM-RoBERTa model proves to be significantly more effective in leveraging multilingual datasets. On Slovenian instances, it achieves scores of only 7 points lower than those achieved on the English subset.

Table 3. Results for the in-domain performance of various models—the models are trained on the training split of the X-GENRE dataset and tested on the test split (X-GENRE-test) from the same dataset.

Model	Micro-F1	Macro-F1	Accuracy
Fine-tuned XLM-RoBERTa	**0.80**	**0.79**	**0.80**
Logistic Regression	0.65	0.67	0.65
SVM	0.66	0.66	0.66
fastText	0.64	0.64	0.64
Naive Bayes	0.56	0.52	0.56
Dummy Classifier	0.13	0.09	0.13

Table 4. The comparison of the performance of various models trained on the X-GENRE training split on English and Slovenian instances. The models are evaluated on the English (EN) and Slovenian (SL) subsets of the X-GENRE test split.

Model	Micro-F1—EN	Micro-F1—SL	Micro-F1 Absolute Difference	Macro-F1—EN	Macro-F1—SL	Macro-F1 Absolute Difference
Fine-tuned XLM-RoBERTa	0.82	0.75	0.07	0.83	0.76	0.07
Logistic Regression	0.71	0.51	0.20	0.73	0.48	0.25
SVM	0.71	0.54	0.17	0.73	0.44	0.29
fastText	0.68	0.55	0.13	0.68	0.43	0.25
Naive Bayes	0.60	0.46	0.14	0.54	0.18	0.36

4.2. Out-of-Domain Performance

Although the performance of the non-neural models in the in-domain experiments appears promising, with accuracy, micro-F1, and macro-F1 scores mostly exceeding 0.64, this does not necessarily indicate that they are suitable to be applied to new data. To assess the models' performance on novel datasets, we conduct out-of-domain experiments, evaluating the same models from the in-domain experiments on a novel EN-GINCO dataset. Table 5 presents the results of these experiments. The performance of all models deteriorates compared to the in-domain experiments. The fine-tuned XLM-RoBERTa model is the only model that demonstrates somewhat satisfactory performance, achieving a micro-F1 score of 0.68 and a macro-F1 score of 0.69. In contrast, the baseline models yield micro-F1 and macro-F1 scores of around 0.50 or lower. McNemar's test confirms that the observed difference between the XLM-RoBERTa model and the second-best performing model, the SVM model, is statistically significant, with a p-value below 0.0005. The drop in performance between the in-domain and out-of-domain results of the XLM-RoBERTa model amounts to 10 to 12 points in both the micro- and macro-F1 scores. In contrast, the Logistic Regression, fastText model, Naive Bayes classifier, and SVM model exhibit a more substantial decline, ranging between 16 and 20 points in the micro-F1 scores and 15 to 23 points in the macro-F1 scores. These findings underscore a crucial insight that evaluating a classifier's performance solely on the dataset it was trained on does not provide adequate information on its robustness or generalizability to new datasets. Consequently,

Mach. Learn. Knowl. Extr. **2023**, *5* 1164

previous studies on automatic genre classification, which predominantly report classifier performance on the same dataset it was trained on, suffer from dataset dependency.

Table 5. The out-of-domain performance of the various models—the models are trained on the training split of the X-GENRE dataset and evaluated on the EN-GINCO test dataset.

Model	Micro-F1	Macro-F1	Accuracy
Fine-tuned XLM-RoBERTa	**0.68**	**0.69**	**0.68**
Logistic Regression	0.49	0.47	0.49
SVM	0.49	0.51	0.49
fastText	0.45	0.41	0.45
Naive Bayes	0.36	0.29	0.36
Dummy Classifier	0.14	0.10	0.14

These findings are consistent with the earlier research conducted by Sharoff et al. (2010) [27], who assessed the robustness of classifiers trained on various genre datasets available at the time. Although the top-performing models achieved accuracy scores exceeding 0.85 in the in-domain scenario, their performance significantly deteriorated in the out-of-domain context, dropping to below 0.40. The authors concluded that the main reason for the low results in the out-of-domain scenario was the inadequate representativeness of the datasets with respect to the actual genres found on the web. However, it is worth noting that the machine learning architecture employed in the study may have also contributed to the observed low performance. Sharoff et al. (2010) [27] used the SVM model, which, similar to our findings, demonstrated limited generalization capabilities. In contrast, the neural model, particularly the fine-tuned XLM-RoBERTa model, exhibits much higher generalization capabilities when trained on the same dataset as the SVM model.

4.3. Zero-Shot Performance of GPT Models

Recent research indicates that for specific text categorization tasks, leveraging recent large instruction-tuned GPT models through prompting can yield comparable performance to the task-specific fine-tuned language models [59,60]. Furthermore, small-scale preliminary experiments using the ChatGPT web interactive interface have shown that the ChatGPT model, which is based on the GPT-3.5 model, outperforms the fine-tuned XLM-RoBERTa model on a sample of the EN-GINCO dataset [62]. Subsequently, we extend these preliminary experiments by including the GPT-3.5 and the GPT-4 models in our comparison of machine learning technologies. Moreover, we use the models via an API instead of the ChatGPT interactive interface. The GPT models were fine-tuned for dialogue and were not specifically trained for the task of automatic genre identification on the X-GENRE dataset. We can be certain of this because the X-GENRE dataset was not published before; therefore, model contamination is not a possibility in this specific case. This is why we consider this a zero-shot scenario, despite providing the models with some contextual information related to the task through the use of prompts.

Table 6 presents the performance results of the GPT-3.5 and GPT-4 models on both the EN-GINCO dataset and the test split of the X-GENRE dataset. The GPT-4 model outperforms the GPT-3.5 model but still exhibits inferior performance compared to the fine-tuned XLM-RoBERTa model. It achieves micro-F1 scores of between 0.65 and 0.7 and macro-F1 scores of between 0.55 and 0.66. However, these results are still impressive, considering that the model was not specifically fine-tuned for the task at hand and that these results were achieved without relying on an extensive training dataset, as is the case with the fine-tuned XLM-RoBERTa model. In terms of the X-GENRE test split, the XLM-RoBERTa model outperforms the GPT-4 model by 10 points in the micro-F1 score and 13 points in the macro-F1 score. The statistical significance of the difference between the models is further supported by the application of McNemar's test, which yields a p-value below the threshold of 0.05, as shown in Table 7. However, it is crucial to acknowledge that comparing

Mach. Learn. Knowl. Extr. **2023**, *5* 1165

the in-domain performance of one model to the zero-shot performance of another model is positively biased toward the model that is tested in the easier, in-domain scenario.

When the models are tested on the EN-GINCO dataset, where the out-of-domain performance of the fine-tuned XLM-RoBERTa model is observed, the discrepancy between the three models is narrower and the statistical analysis using McNemar's test indicates that the difference in the results is not statistically significant, with the observed p-values above the threshold of 0.05. In the micro-F1 score, the XLM-RoBERTa model surpasses the GPT-4 model by only 3 points and the GPT-3.5 model by 5 points. In contrast, the difference is more pronounced in the macro-F1 scores, with the GPT-4 and GPT-3.5 models achieving scores of 14 and 16 points lower than those of the XLM-RoBERTa model, respectively. This suggests that these GPT models are less adept at modeling certain genre categories. These findings provide opportunities for further research, including further prompt engineering experiments that would include more detailed descriptions of genre categories to enhance model performance.

Our study does not confirm the findings of Kuzman et al. (2023) [62], as the GPT-3.5 model, used through the API, did not outperform the fine-tuned XLM-RoBERTa model and achieved lower results compared to when it was used through the ChatGPT interactive interface. However, it is important to note that it is possible that the ChatGPT interface and the API use different model versions, which may be based on different architectures, training data, or other factors that could influence the responses. Additionally, the models used through the web interactive interface and API may need to handle different levels of user load and may have different scaling mechanisms, which could result in differences in their performance. Furthermore, the web interface does not provide any insights into the hyperparameters used for generating a prediction. It is plausible that the models use different hyperparameters, which may have also contributed to the observed differences in the results.

Table 6. Zero-shot performance of the GPT-3.5 and GPT-4 models on the X-GENRE test subset and the EN-GINCO test dataset compared to the performance of the XLM-RoBERTa model, fine-tuned on the training split of the X-GENRE dataset and evaluated in the in-domain (on X-GENRE-test) and out-of-domain (on EN-GINCO) scenarios.

Model	Test Dataset (Evaluation Scenario)	Micro-F1	Macro-F1	Accuracy
Fine-tuned XLM-RoBERTa	EN-GINCO (out-of-domain)	0.68	0.69	0.68
GPT-4	EN-GINCO (zero-shot)	0.65	0.55	0.65
GPT-3.5	EN-GINCO (zero-shot)	0.63	0.53	0.63
Fine-tuned XLM-RoBERTa	X-GENRE-test (in-domain)	0.80	0.79	0.80
GPT-4	X-GENRE-test (zero-shot)	0.70	0.66	0.70
GPT-3.5	X-GENRE-test (zero-shot)	0.65	0.63	0.65

Table 7. The results of McNemar's test for statistical significance for each pair of models evaluated on the EN-GINCO and X-GENRE test sets. Asterisks denote p-values: ** for $p < 0.001$ and * for $p < 0.01$.

Pair of Models	Test Set	Test Statistic	p-Value
XLM-RoBERTa, GPT-3.5	EN-GINCO	2.33	0.127
XLM-RoBERTa, GPT-4	EN-GINCO	0.71	0.399
GPT-3.5, GPT-4	EN-GINCO	0.64	0.423
XLM-RoBERTa, GPT-3.5	X-GENRE-test	49.82	0.000 **
XLM-RoBERTa, GPT-4	X-GENRE-test	26.47	0.000 **
GPT-3.5, GPT-4	X-GENRE-test	7.86	0.005 *

As with the other models in the previous section, we also conduct an evaluation of the capabilities of the GPT-3.5 and the GPT-4 models in multilingual classification. Table 8 presents their performance on the English and Slovenian portions of the X-GENRE test split

Mach. Learn. Knowl. Extr. **2023**, *5* 1166

in comparison to the performance of the XLM-RoBERTa model. As noted earlier, the XLM-RoBERTa model exhibits superior performance, which is expected given that it is trained on the training split of the same dataset, whereas the GPT models are employed in a zero-shot scenario. However, the primary objective of this section of the evaluation is to examine the differences in the models' performance across various languages. The obtained results, presented in Table 8, reveal that all three models exhibit consistent performance between English and Slovenian, with the maximum discrepancy between the two languages being 7 points in the micro- and macro-F1 scores. Interestingly, the GPT-3.5 model exhibits the greatest consistency, with a marginal 1-point variation in both the micro-F1 and macro-F1 scores between its performance in English and Slovenian. These findings demonstrate the potential of the prompting zero-shot classification method in the task of automatic genre identification. This approach proves to be effective in low-resource languages, eliminating the necessity of creating a genre dataset for each language.

Table 8. The comparison of the performance of the fine-tuned XLM-RoBERTa, GPT-3.5, and GPT-4 models on the English and the Slovenian subsets of the X-GENRE test split.

Model	Micro-F1—EN	Micro-F1—SL	Micro-F1 Absolute Difference	Macro-F1—EN	Macro-F1—SL	Macro-F1 Absolute Difference
Fine-tuned XLM-RoBERTa	0.82	0.75	0.07	0.83	0.76	0.07
GPT-4	0.70	0.68	0.02	0.68	0.63	0.05
GPT-3.5	0.65	0.64	**0.01**	0.63	0.62	**0.01**

5. Discussion

In the present study, we investigated the technologies and datasets that can be used for a robust automatic identification of genres, which can be employed for the automatic enrichment of large text collections with genre information. The metadata obtained as a result of this process can provide valuable insights into the quality of text collections, which is a crucial foundation for the development of language technologies. To achieve this objective, we conducted experiments on multiple genre datasets and compared various machine learning technologies that are commonly used for text categorization. In summary, this paper made the following contributions:

- It conducted a controlled comparison of different machine learning approaches for automatic genre identification in both in-domain and out-of-domain scenarios.
- It analyzed the performance of the models on a high-resource language (English) and a low-resource language (Slovenian).
- It provided a freely accessible XLM-RoBERTa-based genre classifier, which enables automatic genre identification in numerous languages (available at https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier (accessed on 6 August 2023)).
- It explored the zero-shot capabilities of recent large GPT Transformer models for automatic genre identification.
- It proposed a unified genre schema, which facilitates the merging of diverse genre datasets, and introduced a new multilingual genre dataset, demonstrating that combining multiple datasets can enhance the model's performance in this task.
- It established a benchmark (available at https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark (accessed on 6 August 2023)) for evaluating the out-of-domain performance of genre classifiers. The benchmark includes the results of our comparison of the models and provides access to the EN-GINCO dataset upon request for any researchers who wish to participate in the benchmarking process.

Mach. Learn. Knowl. Extr. **2023**, *5* 1167

We conducted a comparison of various machine learning architectures in in-domain and out-of-domain scenarios. The in-domain scenario involved training the models on the training split and testing them on the test split that originated from the same dataset, whereas the out-of-domain scenario involved testing the models on a novel dataset. An overview of the results in terms of micro-F1 scores is shown in Figure 3. The experiments revealed that the fine-tuned XLM-RoBERTa model performed the best in the in-domain scenario. In the in-domain scenario, most machine learning models achieved satisfactory results to some extent. However, when these models were tested in an out-of-domain scenario, the reliability of the in-domain results was called into question. The outcomes obtained in the out-of-domain scenario suggest that pre-Transformer-based machine learning models lack the ability to generalize to new datasets. In contrast, the fine-tuned XLM-RoBERTa model demonstrated satisfactory performance in the out-of-domain scenario.

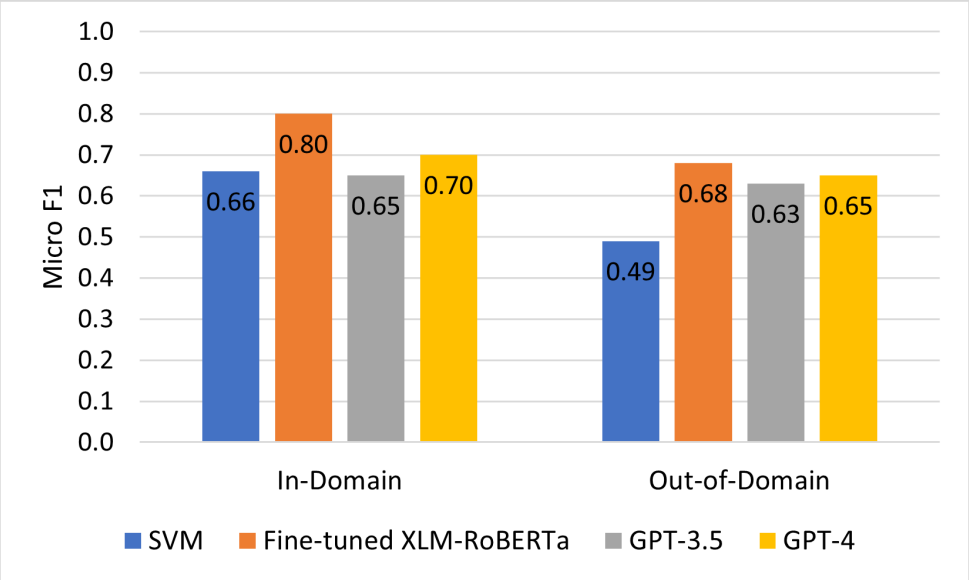


Figure 3. Overview of the performance of the SVM and fine-tuned XLM-RoBERTa model, which were trained on the training split of the X-GENRE dataset and evaluated in in-domain (X-GENRE test split) and out-of-domain (EN-GINCO test dataset) scenarios, and the performance of the GPT-3.5 and GPT-4 models, which were not explicitly trained for the task and were thus used in a zero-shot approach. The figure shows that the fine-tuned Transformer model performed the best in both scenarios. While traditional bag-of-words classification approaches achieved similar performance as the zero-shot prompting approach in the in-domain setting, their performance significantly decreased when applied to novel datasets, whereas the zero-shot approach was, as expected, more robust to domain shifts.

We also included in the comparison the recently developed GPT-3.5 and GPT-4 models, provided by OpenAI, since they have demonstrated promising performance in various text classification tasks [59–61,63], including automatic genre identification [62]. The GPT models were used via prompting, and we considered these experiments to be zero-shot since the models were not explicitly fine-tuned for the task using a labeled dataset. The results of our experiments showed that the fine-tuned XLM-RoBERTa model still had an advantage in terms of performance over the GPT-4 and GPT-3.5 models as long as the fine-tuning data were similar to the testing data. However, on the EN-GINCO dataset, on which none of the models were trained, the difference between them was not statistically significant. The results of these GPT models were impressive, considering that the models were not specifically fine-tuned for the task at hand and that these results were obtained without relying on an extensive training dataset manually annotated with genres. All Transformer-based language models were shown to be capable of identifying genres not

Mach. Learn. Knowl. Extr. **2023**, *5* 1168

only in English but also in Slovenian. These findings illustrate the promising usefulness of the GPT-3.5 and GPT-4 models for automatic genre identification. They could be applied to low-resource languages, without the need for constructing a dedicated genre dataset for each language. Moreover, these results pave the way for exploring the applicability of GPT models to other text classification tasks beyond genre identification, such as sentiment analysis, topic categorization, and hate speech detection.

In conclusion, our experiments demonstrate that fine-tuning the XLM-RoBERTa model on a dataset comprising multiple genre datasets in two languages yields the best results in automatic genre identification and ensures robust performance. However, it is important to note that the performance of fine-tuned BERT-like models is dependent on the availability of a substantial amount of labeled data. Constructing such datasets can be a time-consuming and labor-intensive process, and might be unfeasible for low-resource languages with limited funding. In this regard, newer instruction-tuned GPT models offer a significant advantage, as they achieve high performance without requiring any labeled data. Only smaller manually annotated test datasets are needed to evaluate the performance of these models. On the other hand, the outputs of fine-tuned BERT-like models are more reliable, as they are restricted to a closed set of labels, whereas the outputs of generative GPT-like models are less predictable. To mitigate this disadvantage, careful post-processing of the output of GPT models is necessary. Another disadvantage of instruction-tuned GPT models is that their performance is very sensitive to the content of the prompts. Expertise in prompt engineering is required to elicit the desired responses, which is a non-trivial task. At present, the most effective GPT models are either closed source, such as the OpenAI models, or require high-performance machines to run, such as the Falcon models [3]. Using them for the automatic enrichment of massive text collections would thus incur significant expenses. Furthermore, using the OpenAI GPT models via the API results in considerably slower inference times compared to the local usage of a fine-tuned XLM-RoBERTa model. Therefore, our fine-tuned XLM-RoBERTa classifier, which is freely available online, currently holds a significant advantage in terms of accessibility and speed. It enables the automatic genre annotation of massive web corpora, comprising millions of texts, in just a few hours. Recently, it was employed to enrich Slovenian, Serbian, and Croatian massive monolingual MaCoCu [14] corpora with genre information. These corpora are accessible for querying in freely accessible concordancers (https://www.clarin.si/info/concordances/ (accessed on 6 August 2023)) under the name CLASSLA-web corpora.

Future Work

In future research, we intend to investigate whether the performance of the fine-tuned XLM-RoBERTa model and the GPT-4 model can be further improved. To achieve this, we plan to extend the X-GENRE dataset by (a) incorporating larger parts of the CORE [17] and FTD [28] datasets, which may, however, result in a reduced representation of the Slovenian dataset; and (b) including other languages in the dataset by mapping the X-GENRE schema to the Russian FTD dataset [28], and the Finnish [66], Swedish, and French datasets [65], which use the CORE schema. Furthermore, the performance of the GPT-4 model might also be further improved by further experiments with prompt engineering. We intend to experiment with advanced prompting techniques, such as manual few-shot chain-of-thought prompting [77], or by providing more context to the model by adding descriptions of the genre labels to the prompt. Additionally, further work might analyze the impact of the genre schema on GPT-4 predictions by using different genre schemata in the prompts, including the CORE schema [17], FTD schema [28], and GINCO schema [19]. In addition, we plan to experiment with whether the performance of GPT-like models can be further improved with fine-tuning, benefiting from the Low-Rank Adaptation (LoRA) [78] and QLoRA [79] approaches. These approaches enable the reduction of the memory demands of fine-tuning while maintaining model performance, thereby enabling the fine-tuning of large GPT models on a single GPU.

Mach. Learn. Knowl. Extr. **2023**, *5* 1169

The field of large language models is progressing at a rapid pace. Following the introduction of the GPT-3.5 model in the latter part of 2022, a multitude of new and competitive models have emerged within a short time frame. These include the GPT-4 model [63], the Falcon-40B-Instruct [3] model, and the Llama-2-chat [80] model. Therefore, it is reasonable to expect that the performance of instruction-tuned GPT models in the task of automatic genre identification will continue to improve as new models emerge. Once these models have become available, either through an API or as open source models, we plan to evaluate their performance and incorporate the results into our Automatic Genre Identification Benchmark.

The large language models show very promising performance in the automatic genre identification task, including both out-of-domain and multilingual scenarios. However, it is also important to acknowledge their limitations. One such limitation is their lack of explainability, as their architecture functions as a black box, making it difficult to understand which factors contribute to their prediction of genres. Additionally, the training and inference processes of these models often require significant computational resources, including high-performance hardware and substantial memory, making them less accessible to individuals, researchers, and organizations with resource constraints. Further work is needed to address the challenges posed by the nature of large language models. Furthermore, additional challenges related to the task of automatic genre identification remain to be solved, i.e., performing multi-label classification on hybrid texts and classifying genres to spans of a multi-text document. Although these text categorization challenges are beyond the scope of the current paper, they should be addressed by the research community in the near future.

Author Contributions: Conceptualization, T.K., I.M. and N.L.; Data curation, T.K.; Formal analysis, T.K.; Funding acquisition, N.L.; Investigation, T.K.; Methodology, T.K.; Project administration, N.L.; Resources, T.K.; Software, T.K.; Supervision, I.M. and N.L.; Validation, T.K.; Visualization, T.K.; Writing—original draft, T.K.; Writing—review and editing, I.M. and N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work received funding from the European Union's Connecting Europe Facility 2014–2020—CEF Telecom—under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the authors' views. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research program "Language resources and technologies for Slovene" (P6-0411).

Data Availability Statement: This study includes the following publicly available datasets: the Corpus of Online Registers of English (CORE) [17] dataset, available at https://github.com/TurkuNLP/CORE-corpus (accessed on 30 November 2022); the English Functional Text Dimensions (FTD) [28] dataset, available at https://github.com/ssharoff/genre-keras (accessed on 30 November 2022); and the Slovenian Genre Identification Corpus (GINCO) [19] dataset, published at the CLARIN.SI repository (http://hdl.handle.net/11356/1467 (accessed on 30 November 2022)). Furthermore, we introduce two new genre datasets—the X-GENRE dataset, which consists of instances from the FTD, GINCO, and CORE datasets, and a newly annotated EN-GINCO dataset. These two datasets are available on request from the corresponding author. The data are not publicly available, as we do not own the copyright to the texts included in the datasets. Moreover, we refrain from publishing the datasets to ensure that large language models are not trained using these data. By taking this precautionary measure, we can proceed with evaluating future models using these datasets without the possibility of data leakage from the training dataset. However, we have made the best-performing XLM-RoBERTa model, fine-tuned on the X-GENRE dataset, freely available at the HuggingFace repository (https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier (accessed on 6 August 2023)).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SVMs	Support Vector Machines
NLP	Natural Language Processing
PLM	Pre-Trained Language Model
FTD	Functional Text Dimensions
CORE	Corpus of Online Registers of English
GINCO	Genre Identification Corpus

Appendix A. More Details on Genre Datasets

Appendix A.1. Collection and Annotation Procedure of Existing Datasets

The **CORE** [17] dataset was developed by extracting texts from the "General" part of the Corpus of Global Web-based English (GloWbE) dataset, which was collected by searching for texts on Google based on queries consisting of high-frequency English 3-grams [81]. The dataset was manually annotated via crowd-sourcing with 908 participants, where each text was annotated by four annotators.

The **GINCO** [19] dataset was collected by taking a random sample of 501 instances from each of two Slovenian web corpora from different time periods: the slWaC 2.0 corpus [82] from 2014 and the MaCoCu-sl 1.0 corpus [83] from 2021. The web corpora were created by crawling the Slovenian top-level web domain (.si) and connected websites from common domains, e.g., .com. The texts were manually annotated with genre categories by two expert annotators.

The English **FTD** dataset [28] was collected from two sources: the ukWac web corpus [15], collected in 2006 by crawling the .uk domain, and the Pentaglossal corpus (5 g) [84], which comprises fiction, corporate communication, political debates, TED talks, UN reports, and other texts. The dataset was annotated by Linguistics and Translation MA students, and each text was labeled by at least two annotators.

Appendix A.2. X-GENRE Labels

Table A1. Descriptions of genre labels from the X-GENRE schema, with examples.

Label	Description	Examples
Information/Explanation	An objective text that describes or presents an event, a person, a thing, a concept, etc. Its main purpose is to inform the reader about something.	research article, encyclopedia article, product specification, course materials, biographical story/history
Instruction	An objective text that instructs the readers on how to do something.	how-to texts, recipes, technical support
Legal	An objective formal text that contains legal terms and is clearly structured.	small print, software license, terms and conditions, contracts, law, copyright notices
News	An objective or subjective text that reports on an event, recent at the time of writing or coming in the near future.	news report, sports report, police report, announcement
Opinion/Argumentation	A subjective text in which the authors convey their opinion or narrate their experiences. It includes the promotion of an ideology and other non-commercial causes.	review, blog, editorial, letter to editor, persuasive article or essay, political propaganda

Mach. Learn. Knowl. Extr. **2023**, *5* 1171

Table A1. *Cont.*

Label	Description	Examples
Promotion	A subjective text intended to sell or promote an event, product, or service. It addresses the readers, often trying to convince them to participate in something or buy something.	advertisement, e-shops, promotion of an accommodation, promotion of a company's services, an invitation to an event
Prose/Lyrical	A literary text that consists of paragraphs or verses. A literary text is deemed to have no other practical purpose than to provide pleasure to the reader. Often the author pays attention to the aesthetic appearance of the text. It can be considered as art.	lyrics, poem, prayer, joke, novel, short story
Forum	A text in which people discuss a certain topic in the form of comments.	discussion forum, reader/viewer responses, QA forum
Other	A text that does not fall under any other genre category.	

Appendix A.3. Label Distribution in X-GENRE and EN-GINCO

Table A2. Comparison of the label distribution in the EN-GINCO test set and the X-GENRE dataset.

Label	EN-GINCO	X-GENRE
Information/Explanation	25%	17%
Promotion	22%	16%
Opinion/Argumentation	18%	14%
News	18%	19%
Other	6%	4%
Forum	6%	8%
Instruction	5%	12%
Legal	0%	4%
Prose/Lyrical	0%	6%

Table A2 shows the differences in the label distributions between the X-GENRE and EN-GINCO datasets. Two categories from the X-GENRE dataset are not present in the EN-GINCO test set: *Legal* and *Prose/Lyrical*. The comparison also shows that the EN-GINCO dataset consists of more *Information/Explanation*, *Promotion*, and *Opinion/Argumentation* instances than the X-GENRE dataset, while the *Instruction* category is less represented. As the label distributions differ, training the models on the X-GENRE dataset and evaluating them on the EN-GINCO test set provides valuable insights into their performance in an out-of-distribution scenario.

Appendix B. Model Hyperparameters

We used the following hyperparameters for the models, as presented in Section 3.2:

- Dummy Classifier: stratified strategy.
- Naive Bayes: ComplementNB model with the default hyperparameters.
- Logistic Regression: Penalty set to None, as preliminary experiments showed that disabling regularization improved the results in our case.
- SVM: SVC model with the linear kernel and the regularization parameter C set to 2.
- fastText: The number of epochs was set to 350 and word unigrams were used (wordNgrams set to 1).
- XLM-RoBERTa: We used a learning rate of 1×10^{-5} and a maximum sequence length of 512 for all fine-tuned models, as described in Section 3.2.2. However, the

models differed in the optimum number of epochs, which was determined based on a hyperparameter tuning, tested on the evaluation split. The optimum number of epochs varied based on the training dataset used. The hyperparameter search determined the following numbers of epochs to be optimal for the specific genre dataset: CORE-main—4 epochs; CORE-sub—6 epochs; FTD—10 epochs; GINCO-downcast and X-GENRE—15 epochs; GINCO-reduced—20 epochs.

- GPT-3.5 and GPT-4: We used the GPT-3.5 Turbo model for the GPT-3.5 model and the GPT-4 model. The models were used through the chat completion endpoint. We set the temperature to 0, which ensured that the model was more deterministic and that it output the prediction with the highest probability. To facilitate the parsing of the model's output, we limited the number of tokens in the output (maxtokens) to 2 and defined a line break to be the point where the model stops its completion (the stop hyperparameter).

Appendix C. Instance of a Prompt for the GPT Models

Example of a prompt:

Please classify the following text according to genre (defined by function of the text, author's purpose and form of the text). You can choose from the following classes: News, Legal, Promotion, Opinion, Instruction, Information, Literature, Forum, Other. The text to classify: Shower pods install in no time… <p> 1. Prepare the floor with the waste and the water supply pipes. <p> 2. Attach shower equipment to the shower pod shell running flexible tails (H&C or just C) down back. <p> 3. Move the unit into position connecting water supplies on the way and the waste outlet trap. <p> 4. Having secured the shower pod shell to the building structure doors may now be fitted.

Class:

References

1. Zu Eissen, S.M.; Stein, B. Genre classification of web pages. In *Proceedings of the 27th Annual German Conference in AI, KI 2004, Ulm, Germany, 20–24 September 2004*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 256–269.
2. Vidulin, V.; Luštrek, M.; Gams, M. Using genres to improve search engines. In Proceedings of the 1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing, Borovets, Bulgaria, 30 September 2007; pp. 45–51.
3. Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; Launay, J. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv* **2023**, arXiv:2306.01116.
4. Kuzman, T.; Rupnik, P.; Ljubešić, N. Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. In Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Dubrovnik, Croatia, 5–6 May 2023; pp. 91–103.
5. Orlikowski, W.J.; Yates, J. Genre repertoire: The structuring of communicative practices in organizations. *Adm. Sci. Q.* **1994**, *39*, 541–574. [CrossRef]
6. Joulin, A.; Grave, É.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 427–431.
7. Stubbe, A.; Ringlstetter, C. Recognizing genres. In Proceedings of the Towards a Reference Corpus of Web Genres, Birmingham, UK, 27 July 2007.
8. Finn, A.; Kushmerick, N. Learning to classify documents according to genre. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 1506–1518. [CrossRef]
9. Roussinov, D.; Crowston, K.; Nilan, M.; Kwasnik, B.; Cai, J.; Liu, X. Genre based navigation on the web. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 6 January 2001.
10. Priyatam, P.N.; Iyengar, S.; Perumal, K.; Varma, V. Don't Use a Lot When Little Will Do: Genre Identification Using URLs. *Res. Comput. Sci.* **2013**, *70*, 233–243. [CrossRef]
11. Boese, E.S. Stereotyping the Web: Genre Classification of Web Documents. Ph.D. Thesis, Colorado State University, Fort Collins, CO, USA, 2005.
12. Stein, B.; Eissen, S.M.Z.; Lipka, N. Web genre analysis: Use cases, retrieval models, and implementation issues. In *Genres on the Web*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 167–189.
13. Crowston, K.; Kwaśnik, B.; Rubleske, J. Problems in the use-centered development of a taxonomy of web genres. In *Genres on the Web*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 69–84.

14. Bañón, M.; Esplà-Gomis, M.; Forcada, M.L.; García-Romero, C.; Kuzman, T.; Ljubešić, N.; van Noord, R.; Sempere, L.P.; Ramírez-Sánchez, G.; Rupnik, P.; et al. MaCoCu: Massive collection and curation of monolingual and bilingual data: Focus on under-resourced languages. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, Ghent, Belgium, 1–3 June 2022; pp. 301–302.
15. Baroni, M.; Bernardini, S.; Ferraresi, A.; Zanchetta, E. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Lang. Resour. Eval.* **2009**, *43*, 209–226. [CrossRef]
16. Sharoff, S. In the Garden and in the Jungle. In *Genres on the Web*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 149–166.
17. Egbert, J.; Biber, D.; Davies, M. Developing a bottom-up, user-based method of web register classification. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 1817–1831. [CrossRef]
18. Laippala, V.; Kyllönen, R.; Egbert, J.; Biber, D.; Pyysalo, S. Toward multilingual identification of online registers. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, 30 September–2 October 2019; pp. 292–297.
19. Kuzman, T.; Rupnik, P.; Ljubešić, N. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In Proceedings of the Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 1584–1594.
20. Giesbrecht, E.; Evert, S. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In Proceedings of the Fifth Web as Corpus Workshop, San Sebastián, Spain, 7 September 2009; pp. 27–35.
21. Müller-Eberstein, M.; van der Goot, R.; Plank, B. Genre as Weak Supervision for Cross-lingual Dependency Parsing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 4786–4802.
22. Van der Wees, M.; Bisazza, A.; Monz, C. Evaluation of machine translation performance across multiple genres and languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
23. Agrawal, S.; Sanagavarapu, L.M.; Reddy, Y.R. FACT-Fine grained Assessment of web page CredibiliTy. In Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON), Kochi, India, 17–20 October 2019; pp. 1088–1097.
24. Rehm, G.; Santini, M.; Mehler, A.; Braslavski, P.; Gleim, R.; Stubbe, A.; Symonenko, S.; Tavosanis, M.; Vidulin, V. Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In Proceedings of the LREC, Marrakech, Morocco, 26 May–1 June 2008.
25. Berninger, V.F.; Kim, Y.; Ross, S. Building a document genre corpus: A profile of the KRYS I corpus. In Proceedings of the BCS-IRSG Workshop on Corpus Profiling, London, UK, 18 October 2008; pp. 1–10.
26. Santini, M. Automatic Identification of Genre in Web Pages. Ph.D. Thesis, University of Brighton, Brighton, UK, 2007.
27. Sharoff, S.; Wu, Z.; Markert, K. The Web Library of Babel: Evaluating genre collections. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010.
28. Sharoff, S. Functional text dimensions for the annotation of web corpora. *Corpora* **2018**, *13*, 65–95. [CrossRef]
29. Asheghi, N.R.; Sharoff, S.; Markert, K. Crowdsourcing for web genre annotation. *Lang. Resour. Eval.* **2016**, *50*, 603–641. [CrossRef]
30. Suchomel, V. Genre Annotation of Web Corpora: Scheme and Issues. In *Future Technologies Conference*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 738–754.
31. Sharoff, S. Genre annotation for the web: Text-external and text-internal perspectives. *Regist. Stud.* **2021**, *3*, 1–32. [CrossRef]
32. Kuzman, T.; Ljubešić, N.; Pollak, S. Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In *Jezikovne Tehnologije in Digitalna Humanistika: Zbornik Konference*; Fišer, D., Erjavec, T., Eds.; Institute of Contemporary History: München, Germany, 2022; pp. 100–107.
33. Repo, L.; Skantsi, V.; Rönnqvist, S.; Hellström, S.; Oinonen, M.; Salmela, A.; Biber, D.; Egbert, J.; Pyysalo, S.; Laippala, V. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021, Online, 19–23 April 2021; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2021; pp. 183–191. Available online: https://aclanthology.org/2021.eacl-srw.24.pdf (accessed on 6 August 2023).
34. Lepekhin, M.; Sharoff, S. Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for the Genre Classification Task. In Proceedings of the Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 5974–5982.
35. Rezapour Asheghi, N. Human Annotation and Automatic Detection of Web Genres. Ph.D. Thesis, University of Leeds, Leeds, UK, 2015.
36. Laippala, V.; Luotolahti, J.; Kyröläinen, A.J.; Salakoski, T.; Ginter, F. Creating register sub-corpora for the Finnish Internet Parsebank. In Proceedings of the 21st Nordic Conference on Computational Linguistics, Gothenburg, Sweden, 22–24 May 2017; pp. 152–161.
37. Petrenz, P.; Webber, B. Stable classification of text genres. *Comput. Linguist.* **2011**, *37*, 385–393. [CrossRef]
38. Laippala, V.; Egbert, J.; Biber, D.; Kyröläinen, A.J. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Lang. Resour. Eval.* **2021**, *55*, 757–788. [CrossRef]
39. Pritsos, D.; Stamatatos, E. Open set evaluation of web genre identification. *Lang. Resour. Eval.* **2018**, *52*, 949–968. [CrossRef]
40. Feldman, S.; Marin, M.A.; Ostendorf, M.; Gupta, M.R. Part-of-speech histograms for genre classification of text. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4781–4784.

41. Biber, D.; Egbert, J. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *J. Res. Des. Stat. Linguist. Commun. Sci.* **2015**, *2*, 3–36. [CrossRef]
42. Dewdney, N.; Van Ess-Dykema, C.; MacMillan, R. The form is the substance: Classification of genres in text. In Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, 6–7 July 2001.
43. Lim, C.S.; Lee, K.J.; Kim, G.C. Multiple sets of features for automatic genre classification of web documents. *Inf. Process. Manag.* **2005**, *41*, 1263–1276.
44. Levering, R.; Cutler, M.; Yu, L. Using visual features for fine-grained genre classification of web pages. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, HI, USA, 7–10 January 2008; p. 131.
45. Maeda, A.; Hayashi, Y. Automatic genre classification of Web documents using discriminant analysis for feature selection. In Proceedings of the 2009 Second International Conference on the Applications of Digital Information and Web Technologies, London, UK, 4–6 August 2009; pp. 405–410.
46. Abramson, M.; Aha, D.W. What's in a URL? Genre Classification from URLs. In Proceedings of the Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Canada, 22–26 July 2012.
47. Jebari, C. A pure URL-based genre classification of web pages. In Proceedings of the 2014 25th International Workshop on Database and Expert Systems Applications, Munich, Germany, 1–5 September 2014; pp. 233–237.
48. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv* **2020**, arXiv:2004.03705.
49. Kuzman, T.; Ljubešić, N. Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In *Proceedings of the Odkrivanje Znanja in Podatkovna Skladišča—SiKDD, Ljubljana, Slovenia, 10 October 2022*; Mladenić, D., Grobelnik, M., Eds.; Institut "Jožef Stefan" : Ljubljana, Slovenia, 2022.
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017; pp. 6000–6010.
51. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 6 August 2023).
52. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, USA, 2–7 June 2019; pp. 4171–4186.
53. Laippala, V.; Rönnqvist, S.; Oinonen, M.; Kyröläinen, A.J.; Salmela, A.; Biber, D.; Egbert, J.; Pyysalo, S. Register identification from the unrestricted open Web using the Corpus of Online Registers of English. *Lang. Resour. Eval.* **2022**, *57*, 1045–1079. [CrossRef]
54. Rönnqvist, S.; Skantsi, V.; Oinonen, M.; Laippala, V. Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Iceland, 31 May–2 June 2021; pp. 157–165.
55. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
56. OpenAI. ChatGPT General FAQ. 2023. Available online: https://help.openai.com/en/articles/6783457-chatgpt-general-faq (accessed on 3 March 2023).
57. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017; pp. 4302–4310.
58. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv* **2023**, arXiv:2302.06476.
59. Zhang, B.; Ding, D.; Jing, L. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv* **2022**, arXiv:2212.14548.
60. Huang, F.; Kwak, H.; An, J. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *arXiv* **2023**, arXiv:2302.07736.
61. Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y.J.; Afify, M.; Awadalla, H.H. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv* **2023**, arXiv:2302.09210.
62. Kuzman, T.; Ljubešić, N.; Mozetič, I. ChatGPT: Beginning of an End of Manual Annotation? Use Case of Automatic Genre Identification. *arXiv* **2023**, arXiv:2303.03953.
63. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
64. Santini, M. Cross-testing a genre classification model for the web. In *Genres on the Web*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 87–128.
65. Laippala, V.; Rönnqvist, S.; Hellström, S.; Luotolahti, J.; Repo, L.; Salmela, A.; Skantsi, V.; Pyysalo, S. From web crawl to clean register-annotated corpora. In Proceedings of the 12th Web as Corpus Workshop, Marseille, France, 11–16 May 2020; pp. 14–22.

66. Skantsi, V.; Laippala, V. Analyzing the unrestricted web: The Finnish corpus of online registers. *Nord. J. Linguist.* **2023**, 1–31. Available online: https://www.cambridge.org/core/journals/nordic-journal-of-linguistics/article/analyzing-the-unrestricted-web-the-finnish-corpus-of-online-registers/BDCA0FE03ABD9087CC5652533880C8C0 (accessed on 6 August 2023) [CrossRef]
67. Jakubíček, M.; Kilgarriff, A.; Kovář, V.; Rychlỳ, P.; Suchomel, V. The TenTen corpus family. In Proceedings of the 7th International Corpus Linguistics Conference CL, Lancaster, UK, 23–26 July 2013 ; pp. 125–127.
68. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
69. Rennie, J.D.; Shih, L.; Teevan, J.; Karger, D.R. Tackling the poor assumptions of Naive Bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 616–623.
70. Byrd, R.H.; Lu, P.; Nocedal, J.; Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208. [CrossRef]
71. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155.
72. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
73. Asheghi, N.R.; Markert, K.; Sharoff, S. Semi-supervised graph-based genre classification for web pages. In Proceedings of the TextGraphs-9: The Workshop on Graph-Based Methods for Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 39–47.
74. Santini, S.M. Common criteria for genre classification: Annotation and granularity. In Proceedings of the Workshop on Text-Based Information Retrieval (TIR-06), Riva del Garda, Italy, 29 August 2006. Available online: https://ceur-ws.org/Vol-205/paper9.pdf (accessed on 6 August 2023).
75. Everitt, B.S. *The Analysis of Contingency Tables*; CRC Press: Boca Raton, FL, USA, 1992.
76. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef]
77. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35, New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837.
78. Hu, E.J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Online, 3–7 May 2021. Available online: https://openreview.net/pdf?id=nZeVKeeFYf9 (accessed on 6 August 2023).
79. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv* **2023**, arXiv:2305.14314.
80. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
81. Davies, M.; Fuchs, R. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *Engl. World-Wide* **2015**, *36*, 1–28. [CrossRef]
82. Erjavec, T.; Ljubešić, N. The slWaC 2.0 corpus of the Slovene web. *Jezikovne Tehnol. Zb.* **2014**, *17*, 50–55. Available online: https://nl.ijs.si/isjt14/proceedings/isjt2014_08.pdf (accessed on 6 August 2023).
83. Bañón, M.; Esplà-Gomis, M.; Forcada, M.L.; García-Romero, C.; Kuzman, T.; Ljubešić, N.; van Noord, R.; Pla Sempere, L.; Ramírez-Sánchez, G.; Rupnik, P.; et al. Slovene Web Corpus MaCoCu-sl 1.0. 2022. Slovenian Language Resource Repository CLARIN.SI. Available online: http://hdl.handle.net/11356/1517 (accessed on 6 August 2023).
84. Forsyth, R.S.; Sharoff, S. Document dissimilarity within and across languages: A benchmarking study. *Lit. Linguist. Comput.* **2014**, *29*, 6–22. [CrossRef]

Chapter 4

State of the Art for Automatic Genre Identification

Having created the multilingual training and test genre dataset, X-GENRE – which we have shown to be well-suited for training genre classifiers (see Section 3.2.2) – we now move to the next step: experimenting with various text classification methods to identify which offers the most robust generalization across datasets and languages. To this end, we train a range of neural and non-neural models on the Slovenian-English X-GENRE dataset, and evaluate them on the test split of the X-GENRE dataset as well as on newly developed test datasets in multiple languages.

The creation of new test datasets in English and ten other European languages, presented in Section 4.1.1, fulfills the second part of Goal *G3*, which aimed to construct separate test datasets for a range of European languages. Building on these datasets, we establish a publicly available benchmark, addressing Goal *G4* that focuses on enabling cross-dataset and cross-lingual evaluation of machine learning approaches for automatic genre identification. This benchmark fills a current gap by providing reference test datasets for the task, allowing for consistent and comparative analyses of genre classification methods.

The models are assessed in the following scenarios: an in-domain (in-dataset) scenario, where the model is trained on a portion of the same dataset from which the test dataset originates, and a cross-dataset scenario, where the model is tested on a dataset developed independently of the training data. We further extend the second scenario to a zero-shot cross-lingual setting, in which the model is evaluated on a test dataset in a language it has not seen during fine-tuning. Additionally, we experiment with a very recent approach to text classification – employing decoder-only instruction-tuned Transformer-based large language models (LLMs) guided by instructions in the form of prompts –, which is an additional form of zero-shot classification, as these models were not specifically trained on a genre dataset.

These evaluations enable us to identify the machine learning methodology that provides the best-performing genre classifier, with which we fulfill Goal *G5*: "Develop a robust multilingual genre classifier capable of generalizing across datasets and European languages by exploring various machine learning approaches, including deep neural language models." Our benchmarking experiments also address Hypothesis *H2* "*Fine-tuned deep neural Transformer-based language models outperform other machine learning methods in the task of automatic genre identification in the Slovenian language.*". The results in Section 4.2 provide evidence consistent with this hypothesis. We additionally assess Hypothesis *H3* "*A reliable zero-shot cross-lingual performance in automatic genre identification across numerous European languages is possible by leveraging the cross-lingual capabilities of massively*

multilingual pre-trained Transformer models.", and we find suggestive evidence in its favor in Section 4.4.

This chapter is structured as follows. In Section 4.1, we detail the methodology and present the newly developed test datasets designed for benchmarking classification methods on automatic genre identification. The results of experiments assessing in-dataset and cross-dataset performance on Slovenian and English are presented in Sections 4.2 and 4.3. Finally, we extend the evaluation to examine the zero-shot cross-lingual performance in Section 4.4 and zero-shot performance of LLMs in Section 4.5. Further details on the experiments, including detailed descriptions of the evaluated methods, are provided in the papers "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4, "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al., In submission) in Section 4.6, and "*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?*" (Kuzman Pungeršek, Rupnik, Porupski, et al., 2026) in Section 4.7.

4.1 Test Datasets and Methodology

To evaluate various classification methods on the task of automatic genre identification, we train the models on the training split of the Slovenian-English X-GENRE dataset (Kuzman & Ljubešić, 2024a; Kuzman, Mozetič, et al., 2023), presented in Section 3.2.2. The models are then evaluated in three scenarios:

- In-Dataset Setup (Section 4.2): Models are evaluated on the test split (X-GENRE-test) of the same dataset from which the training split was drawn. Although the training and test splits contain similar types of instances, there is no overlap between them. As the test set includes both Slovenian and English, it also enables a direct comparison of model performance across the two languages.

- Cross-Dataset Setup (Section 4.3): Models are evaluated on an independently developed English test dataset, the EN-GINCO dataset (Kuzman, Mozetič, et al., 2023), allowing us to assess generalization across datasets in the same language.

- Cross-Lingual Cross-Dataset Setup (Section 4.4): The best-performing model from the previous two setups is evaluated on a newly created X-GINCO test dataset (Kuzman Pungeršek et al., In submission) comprising ten European languages, developed independently of the training data. Except for the evaluation on the Slovenian portion of the test dataset, this setup represents a zero-shot scenario, as the model is tested on languages it has not seen during training. This setup allows us to assess the model's ability to generalize across both datasets and languages, which we regard as robustness.

In addition, we expand our investigation by incorporating a novel approach to text classification, employing decoder-only instruction-tuned Transformer-based large language models (LLMs) guided by instructions in the form of prompts. As these models were not specifically trained on a genre dataset, we regard their performance as zero-shot (Section 4.5).

We compare the following machine learning methods that have been commonly used for text classification tasks (for details on their implementation and hyperparameters, see the papers by Kuzman, Mozetič, et al. (2023) in Section 3.2.4 and Kuzman Pungeršek, Rupnik, Porupski, et al. (2026) in Section 4.7):

- traditional non-neural machine learning methods, namely, a dummy classifier using a stratified strategy, which provides a simple baseline; and the Naive Bayes classifier, logistic regression classifier, and a support vector machine (SVM) with a linear kernel, which use the TF-IDF (term frequency–inverse document frequency) text representations;

- shallow neural models, namely, the fastText model (Joulin et al., 2017);

- encoder-only Transformer-based pretrained BERT-like language models, namely the base-sized multilingual XLM-RoBERTa model (Conneau et al., 2020) which has been shown to be the most suitable BERT-like model for the task of automatic genre identification (Kuzman, Ljubešić, & Pollak, 2022; Repo et al., 2021; Rönnqvist et al., 2021);

- instruction-tuned decoder-only large language models (LLMs), which have shown very competitive performance in various classification tasks (Achiam et al., 2023; Hendy et al., 2023; Huang et al., 2023; Zhang et al., 2022). Specifically, we evaluate a selection of open-source and closed-source models: (a) closed-source GPT-3.5-Turbo (`gpt-3.5-turbo-0125`) (OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024), GPT-5 (`gpt-5-2025-08-07`) (OpenAI, 2025), Gemini 2.5 Flash (Comanici et al., 2025) and Mistral Medium 3.1 (`mistral-medium-2508`) (Mistral AI, 2025), and (b) open-source LLaMA 3.3 (Meta, 2024), Gemma 3 (Gemma Team et al., 2025), Qwen 3 (Yang et al., 2025), and DeepSeek-R1-Distill (`DeepSeek-R1-Distill-Qwen-14B`) (Guo et al., 2025). Details on the models, their implementation and the prompt are provided in the paper by Kuzman Pungeršek, Rupnik, Porupski, et al. (2026), enclosed in Section 4.7.

The model performance is evaluated using two standard metrics for text classification tasks (Kowsari et al., 2019): micro-F1 and macro-F1 scores. The micro-F1 score aggregates precision and recall across all instances, and thus reflects overall classification performance. However, as this metric is more influenced by the most frequent labels, we use macro-F1 as the primary evaluation metric. Macro-F1 is calculated as the arithmetic mean of the F1 scores for each class, providing a balanced view of performance across all labels regardless of their frequency. This makes it a more reliable indicator of a classifier's real-world applicability, where it is important that the model provides high performance across all categories.

4.1.1 Test Datasets

We develop two novel test datasets to enable the evaluation of genre classifiers in cross-dataset and cross-lingual scenarios: the English EN-GINCO dataset and the multilingual X-GINCO dataset.

The English EN-GINCO test dataset was developed by sampling 272 random texts from the English web corpus enTenTen20 (Jakubíček et al., 2013). The texts were then manually annotated by two expert annotators who participated in the annotation of the GINCO dataset (see Section 3.1.2). For the purposes of model evaluation, the GINCO labels were mapped to the X-GENRE schema. More details on the dataset are provided in the paper by Kuzman, Mozetič, et al. (2023), enclosed in Section 3.2.4.

The multilingual X-GINCO test dataset was constructed in a more controlled manner than the EN-GINCO test dataset. It is based on the MaCoCu-Genre web corpora (Kuzman & Ljubešić, 2024b), which were automatically annotated with genre labels using an XLM-RoBERTa classifier fine-tuned on the X-GENRE dataset. To develop the test dataset,

we randomly selected ten instances for each of eight X-GENRE labels from each of ten automatically-annotated MaCoCu-Genre corpora. Instances labeled as *Other* or assigned low classifier confidence scores – representing around 7% of all instances in the MaCoCu-Genre corpora – were not included in the test dataset to ensure a clean and reliable evaluation sample. The selected instances were then manually annotated with X-GENRE labels by an annotator who was previously involved in the GINCO and EN-GINCO annotation campaigns. To avoid annotation bias, the automatic genre predictions used for sampling were hidden from the annotator. The final X-GINCO test dataset consists of 790 instances, balanced across the eight genre labels (*Information/Explanation, Instruction, Legal, News, Opinion/Argumentation, Promotion, Prose/Lyrical, Forum*). It comprises ten European languages, equally represented in the dataset, namely Albanian, Catalan, Croatian, Greek, Icelandic, Macedonian, Maltese, Slovenian, Turkish, and Ukrainian. These languages span nine branches across three major language families (Indo-European, Turkic, and Afro-Asiatic) and use three different scripts: Latin, Greek, and Cyrillic. This diversity enables a robust evaluation of a genre classifier not only across languages but also across writing systems. Further details on the dataset are provided in the paper by Kuzman Pungeršek et al. (In submission) in Section 4.6.

Importantly, the test datasets are available from the authors upon request but have not been published online to prevent their inclusion in the training data of large language models (LLMs). This precaution ensures that LLMs can be evaluated using these datasets without the risk of data leakage from pretraining. Instead of publishing them online, the test datasets have been integrated into the AGILE Benchmark (Automatic Genre Identification Benchmark) on GitHub,[1] which provides a framework for evaluating genre classification across multiple languages. With this, we fulfill Goal *G4* "Establish a benchmark for cross-dataset and cross-lingual assessment of machine learning approaches in automatic genre identification, based on newly developed test datasets in English, Slovenian, and various other European languages."

4.2 In-Dataset Performance

To determine the most effective machine learning approach for automatic genre identification, we begin by evaluating several methods in the simplest setting – the in-dataset scenario. In this setup, models are trained on the training split from the X-GENRE dataset (approximately 1,800 instances) and tested on the corresponding test split (around 600 instances) from the same dataset. Both splits include instances in Slovenian and English language.

As shown in Table 4.1, the best performance in the in-dataset setting is achieved by a fine-tuned BERT-like Transformer model. Specifically, the fine-tuned XLM-RoBERTa model (Conneau et al., 2020) achieves a micro-F1 score of 0.80 and a macro-F1 score of 0.79, meeting the high reliability threshold of 0.80 proposed in Section 1.3. This performance is consistent with results reported in previous studies on automatic genre identification, where the fine-tuned XLM-RoBERTa achieved micro-F1 scores ranging from 0.72 to 0.84 in similar in-dataset classification setups (Repo et al., 2021; Rönnqvist et al., 2021). In contrast, non-neural machine learning approaches – logistic regression, support vector machine (SVM), and a Naive Bayes classifier – as well as the shallow neural fastText model, perform notably worse.

Based on prior work (see Section 2.3) and human performance on this task, measured by inter-annotator agreement (see Section 3.1.2), we consider macro-F1 scores of around 0.70 as the minimum threshold for acceptable performance of a genre classifier. The SVM,

[1] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

Table 4.1: In-dataset performance of different machine learning approaches on the test split of the X-GENRE dataset, reported for the full test split as well as for the Slovenian (SL) and English (EN) subsets. Models are ranked by their macro-F1 scores achieved on the entire test set.

Model	Micro-F1	Macro-F1	Macro-F1 (EN)	Macro-F1 (SL)
Fine-tuned XLM-RoBERTa	**0.80**	**0.79**	0.83	0.76
Logistic regression	0.65	0.67	0.73	0.48
Support vector machine	0.66	0.66	0.73	0.44
fastText	0.64	0.64	0.68	0.43
Naive Bayes classifier	0.56	0.52	0.54	0.18
Dummy classifier	0.13	0.09	0.10	0.08

fastText and logistic regression classifiers lag behind the fine-tuned BERT-like model significantly, but still provide results that are useful to some extent, with the logistic regression classifier and the SVM surpassing the 0.70 macro-F1 threshold on at least the English portion of the test set. In contrast, the Naive Bayes classifier performs significantly worse, with micro-F1 and macro-F1 scores slightly above 0.50, indicating limited generalization even within the same dataset.

Since the X-GENRE test set contains instances in both English and Slovenian, it also enables a comparison of model performance across these two languages. As shown in Table 4.1, all models perform worse on the Slovenian portion of the dataset. The drop in macro-F1 scores on Slovenian, compared to English, ranges from 7 points for the fine-tuned XLM-RoBERTa model to 36 points for the Naive Bayes classifier. Despite this decline, the fine-tuned XLM-RoBERTa still performs well on Slovenian, achieving a macro-F1 score of 0.76. In contrast, the remaining classifiers are shown to be ineffective in this multilingual scenario, with macro-F1 scores falling below 0.50. These findings support Hypothesis *H2*, demonstrating that fine-tuned Transformer-based BERT-like language models outperform other machine learning approaches in automatic genre identification for Slovenian.

4.3 Cross-Dataset Performance

Previous studies have emphasized the importance of evaluating genre classifiers in cross-dataset settings, as in-dataset performance may not accurately reflect the effectiveness of the model in real-world scenarios. Sharoff et al. (2010) demonstrated that although SVM models performed well within the genre dataset they were trained on, their accuracy dropped significantly when tested on other genre datasets. This indicates a strong dependence on the training data and an inability of the model to generalize beyond a single dataset. Therefore, cross-dataset evaluation is essential for assessing the robustness of a genre classifier.

Despite providing worse results than the fine-tuned BERT-like model in the in-dataset setup, most machine learning methods were shown to provide results that are useful to some extent, with macro-F1 and micro-F1 scores above 0.64. However, when these models are tested in the out-of-domain scenario on the English EN-GINCO test dataset, their generalization abilities are called into question. As shown in Table 4.2, all pre-Transformer-based models yield micro-F1 and macro-F1 scores of around 0.50 or lower. This indicates that these models are not suitable for building a robust genre classifier.

The fine-tuned XLM-RoBERTa model also shows a notable decline in performance on the EN-GINCO test dataset, with both micro-F1 and macro-F1 scores dropping by about

Table 4.2: The out-of-domain performance of the various models on the English EN-GINCO test dataset. The models are sorted based on the macro-F1 scores.

Model	Micro-F1	Macro-F1	Accuracy
Fine-tuned XLM-RoBERTa	**0.68**	**0.69**	**0.68**
Support vector machine	0.49	0.51	0.49
Logistic regression	0.49	0.47	0.49
fastText	0.45	0.41	0.45
Naive Bayes classifier	0.36	0.29	0.36
Dummy classifier	0.14	0.10	0.14

10 points compared to the in-dataset scenario. Nevertheless, its macro-F1 score remains close to the 0.70 threshold for minimum acceptable performance and is still significantly higher than that of all other models.

4.4 Zero-Shot Cross-Lingual Performance

Manual annotation of training data for developing a genre classifier in a new language requires substantial time, human effort, and financial resources. Therefore, a classifier capable of achieving high performance on languages without manually-annotated data would be highly valuable. Motivated by this, in this section, we evaluate the best-performing model from the in-dataset and cross-dataset experiments on additional languages on which the model has not been trained. This section is based on the main findings from the paper "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al., In submission), enclosed in Section 4.6.

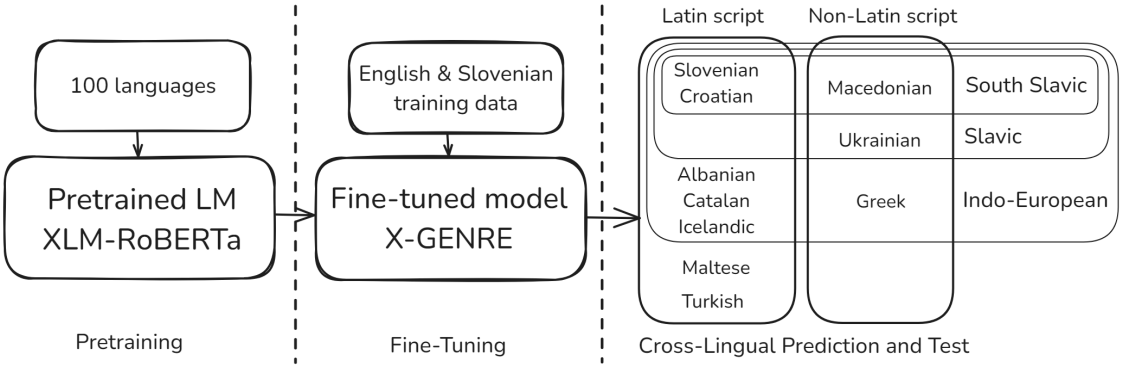


Figure 4.1: Models, languages and tasks involved in the evaluation of the cross-lingual performance of the XLM-RoBERTa-based genre classifier.

Figure 4.1 provides more details on the languages included in the zero-shot cross-lingual experiments. We evaluate the cross-lingual performance of the fine-tuned XLM-RoBERTa model, which achieved the best results in the in-dataset and cross-dataset evaluations (see Sections 4.2 and 4.3). The base XLM-RoBERTa model (Conneau et al., 2020) was pre-trained on 100 languages, covering all test languages in our cross-lingual evaluation except Maltese. For genre classification, the model was fine-tuned on the X-GENRE training set, which includes texts in English and Slovenian (see Section 3.2.2 for details). Evaluation is conducted on the X-GINCO test set, which contains texts in ten languages with varying

degrees of linguistic relatedness to English and Slovenian and written in different scripts: Albanian (sq), Catalan (ca), Croatian (hr), Greek (el), Icelandic (is), Macedonian (mk), Maltese (mt), Slovenian (sl), Turkish (tr), and Ukrainian (uk).

Table 4.3: Per-label performance in terms of F1 scores for each language-specific test set from the X-GINCO dataset, and per-language performance in terms of macro-F1. The penultimate column shows the average F1 score (avg) for each label across all test sets, excluding the scores for the Maltese test set. The last column shows the results for the Maltese test set which is regarded as an outlier. The labels *Information/Explanation* and *Opinion/Argumentation* are abbreviated as *Information* and *Opinion*.

Genre label	ca	el	hr	is	mk	sl	sq	tr	uk	Avg	mt
News	0.82	0.90	0.95	0.73	0.91	0.90	0.89	0.95	1.00	0.89	0.69
Opinion	0.84	0.87	0.78	0.82	0.78	0.82	0.67	0.82	0.91	0.81	0.33
Instruction	0.75	0.71	0.75	0.78	1.00	1.00	0.95	0.90	0.95	0.86	0.69
Information	0.72	0.70	0.95	0.50	0.84	0.90	0.80	0.82	1.00	0.80	0.52
Promotion	0.78	0.62	0.87	0.75	0.95	1.00	0.95	0.86	0.78	0.84	0.82
Forum	0.84	0.95	0.91	0.95	1.00	1.00	0.78	0.89	0.95	0.91	0.18
Prose/Lyrical	0.91	1.00	0.86	1.00	0.95	0.91	0.86	0.95	1.00	0.93	0.18
Legal	0.95	1.00	1.00	0.84	0.95	0.95	0.95	1.00	1.00	0.96	/
Macro-F1	0.83	0.84	0.88	0.80	0.92	0.94	0.85	0.90	0.95	0.87	0.49

As shown in Table 4.3, the fine-tuned XLM-RoBERTa model is highly capable of generalizing across languages in a zero-shot cross-lingual scenario. With the exception of Maltese, the model achieves macro-F1 scores ranging from 0.80 to 0.95. The observed performance is consistent with Hypothesis *H3* "*A reliable zero-shot cross-lingual performance in automatic genre identification across numerous European languages is possible by leveraging the cross-lingual capabilities of massively multilingual pre-trained Transformer models.*". The highest macro-F1 scores are achieved on the Ukrainian, Slovenian, and Macedonian test datasets, with macro-F1 scores exceeding 0.90.

The model's performance in this setting surpasses its results in both the in-dataset and cross-dataset scenarios. This improvement is likely due to the more curated construction of the X-GINCO dataset. In previous evaluations, the model was tested across all nine genre labels, including infrequent and less clearly defined category *Other*. In contrast, instances pre-annotated with a lower prediction confidence or with the label *Other* were excluded from the X-GINCO test dataset prior to manual annotation, resulting in a cleaner dataset with fewer ambiguous instances. This choice is based on the intended real-world application of the genre classifier, where it would be deployed on large-scale text collections and where marking the rare unclear cases with the label *Mix* would ensure an even higher reliability of the remaining annotations. This is why the model is evaluated on clear instances, which constitute the vast majority of web corpora – over 90% of texts –, according to related studies (see paper by Kuzman Pungeršek et al. (In submission) in Section 4.6, and paper by Ljubešić and Kuzman (2024)).

The results in Table 4.3 show that the XLM-RoBERTa classifier underperforms on the Maltese test set, with a macro-F1 score below 0.50. This lower performance is likely due to Maltese not being included in the XLM-RoBERTa pretraining dataset, as observed in previous natural language processing tasks (Chau et al., 2020; de Vries et al., 2022; Ebrahimi & Kann, 2021; Muller et al., 2021). In contrast, the model achieves remarkably high performance even on test sets using non-Latin scripts and on languages that are not closely related to the fine-tuning languages (see Figure 4.1). These results suggest that the

genre classifier has strong potential for use across all languages that have been included in the XLM-RoBERTa pretraining data.

4.5 Zero-Shot Performance of Large Language Models

Until recently, fine-tuned BERT-like models achieved state-of-the-art performance on text classification tasks, and our findings confirm that this also holds for automatic genre identification, as shown throughout this chapter. However, these models require fine-tuning on training datasets developed through time-consuming and costly manual annotation campaigns. In contrast, instruction-tuned decoder-only Transformer models, commonly referred to as large language models (LLMs), have recently demonstrated strong performance across various natural language processing tasks, even in zero-shot prompting scenarios that require no training data (Hendy et al., 2023; Huang et al., 2023; Ljubešić, Galant, et al., 2024; Zhang et al., 2022). In this section, we investigate the effectiveness of the recently introduced zero-shot prompting approach using LLMs for automatic genre identification.

We evaluate a selection of open-source and closed-source LLMs on the English EN-GINCO and multilingual X-GINCO test datasets introduced in Section 4.1.1. To enable a fair comparison between the two datasets, instances labeled as *Other* are removed from the EN-GINCO test dataset, as this category is not included in the more curated X-GINCO dataset. Models are used in a zero-shot prompting setup, meaning that they receive only a task description and label definitions, without any exposure to manually-annotated data. The methodology, the prompt, and the models are described in detail in Kuzman Pungeršek, Rupnik, Porupski, et al. (2026) (see Section 4.7). While that study focuses on English and South Slavic languages, we extend the evaluation to all languages represented in the EN-GINCO and X-GINCO test datasets, using results from the AGILE benchmark.[2]

As shown in Figure 4.2, both open-source and closed-source large language models (LLMs) achieve strong zero-shot performance across all languages except Maltese. Most LLMs surpass the estimated minimum threshold for acceptable performance, defined as a macro-F1 score of approximately 0.70 based on model performance achieved in prior work (Repo et al. (2021), Rönnqvist et al. (2021), and Sharoff (2021); see Table 2.3 in Section 2.3) and human performance on this task. The top-performing models are the closed-source GPT-5 (OpenAI, 2025), GPT-4o (OpenAI, 2024), and Gemini 2.5 Flash (Comanici et al., 2025), with average macro-F1 scores across languages exceeding 0.76 and peak scores reaching up to 0.87, surpassing the high reliability threshold of 0.80 defined in Section 1.3. Although GPT-5 is newer and reportedly more powerful, it does not consistently outperform GPT-4o across all datasets. Open-source models, particularly LLaMA 3.3 (Meta, 2024) and Gemma 3 (Gemma Team et al., 2025), achieve average performance that is more or less comparable to their closed-source counterparts and even outperform them in some languages. The weakest results are observed with the GPT-3.5-Turbo model (OpenAI, 2023), which is an older version of the GPT models, and the DeepSeek-R1-Distill model (Guo et al., 2025), likely due to its smaller size (14 billion parameters). The performance gap between the GPT-3.5-Turbo model (2023) and the GPT-5 model (2025) highlights the rapid pace of progress in LLM development.

As in the cross-lingual experiments discussed in the previous section, Maltese again emerges as an outlier in the LLM evaluations, proving particularly challenging for most models. This is likely due to its limited presence in the pretraining data of current LLMs. However, further empirical analysis is required to assess this explanation. Despite this,

[2]https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

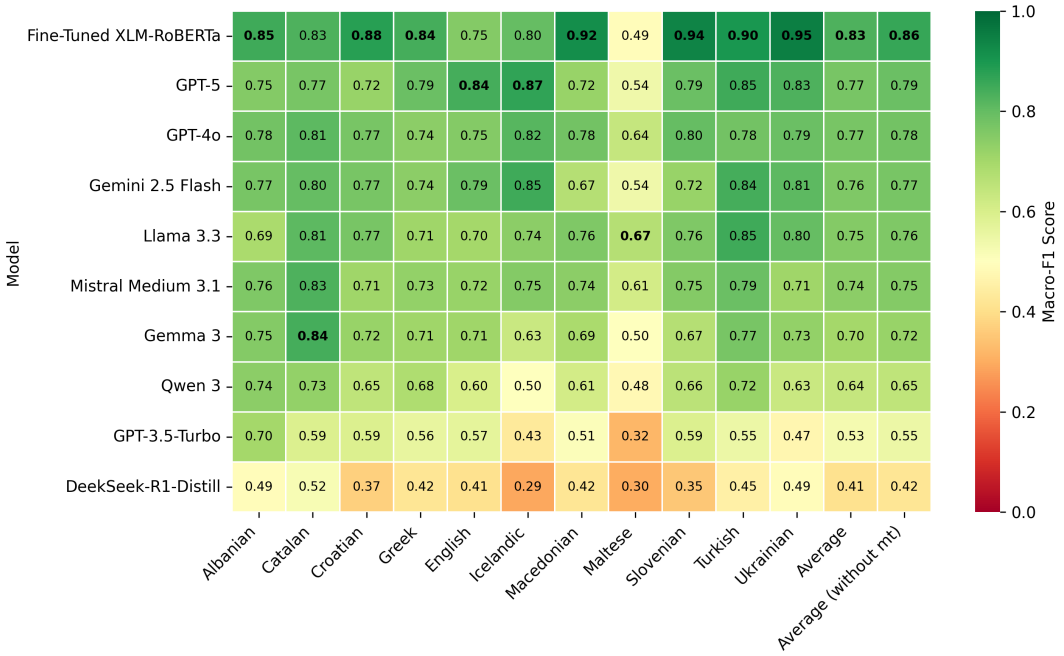


Figure 4.2: Macro-F1 performance on the multilingual X-GINCO and English EN-GINCO test datasets across decoder-only large language models used in a zero-shot prompting scenario, compared to the performance of an XLM-RoBERTa model, fine-tuned on the X-GENRE training dataset. The final column shows the average performance across languages, excluding Maltese as it is an outlier. Results are ordered by the average performance shown in the penultimate column.

some models, namely, the closed-source GPT-4o and the open-source LLaMA 3.3, achieve relatively strong results on the Maltese test set, with macro-F1 scores of 0.64 and 0.67, respectively. These scores approach the threshold for minimum acceptable performance. As LLM development continues at a rapid pace, and if model developers prioritize including underrepresented languages like Maltese in pretraining data, further improvements in performance on this language can be expected.

To enable comparison with the fine-tuned XLM-RoBERTa model, identified in previous sections as the best-performing model, we include its results in Figure 4.2. Specifically, we add its performance on the X-GINCO dataset as reported in Table 4.3, and we recalculate its performance on the EN-GINCO dataset after removing instances labeled as *Other*. As shown in Figure 4.2, the XLM-RoBERTa model, fine-tuned on the X-GENRE training dataset, outperforms large language models in most languages. This shows that fine-tuning on training instances that are relatively similar to the test data (both originating from the web) and that are annotated using consistent guidelines allows the model to better align with the task domain and label definitions. This gives it an advantage over zero-shot prompting approaches, where models rely solely on general knowledge of genres and their definitions without being exposed to concrete examples. However, it is important to note that the strong performance of the fine-tuned XLM-RoBERTa model is also influenced by the curated nature of the X-GINCO test dataset. More challenging instances were filtered out prior to manual annotation based on the prediction confidence of a fine-tuned XLM-RoBERTa classifier, due to which this model reaches high performance on this dataset more easily. Despite this limitation, the X-GINCO dataset remains a valuable benchmark for evaluating LLM performance. In some languages, such as Catalan, Icelandic, and Maltese,

the best-performing LLMs already surpass the fine-tuned XLM-RoBERTa model.

These results demonstrate that large language models, used in a zero-shot prompting setup, offer performance that is highly competitive with fine-tuned BERT-like Transformer-based models. However, we argue that the latter remain the more appropriate choice for automatic genre identification, particularly in large-scale annotation scenarios. LLMs are significantly more computationally expensive and notably slower in terms of inference speed (see Kuzman Pungeršek, Rupnik, Porupski, et al. (2026) in Section 4.7). In addition, BERT-like models, such as XLM-RoBERTa, provide greater control over the annotation process, as they output probabilities for each predefined class. This allows for the calculation of prediction confidence and the application of confidence thresholds to the automatically-annotated data, with which we can further improve annotation reliability. In contrast, LLMs are not constrained to a fixed label set and are prone to producing invalid labels. Our experiments have shown that the models occasionally deviated from the defined label set, with some of the models producing non-existing labels for up to 4% of test instances. These criteria – output reliability, inference speed, and computational cost – are especially important when the goal is to annotate text collections comprising millions of texts, such as massive web corpora. In these cases with extensive data to be processed, fine-tuned BERT-like models remain the most practical and effective solution.

4.6 Related Paper: Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages

Natural Language Processing (2025), 1–25
doi:10.1017/xxxxx

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages

Taja Kuzman Pungeršek, Igor Mozetič, and Nikola Ljubešić

Department of Knowledge Technologies, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
taja.kuzman@ijs.si
Department of Knowledge Technologies, Jožef Stefan Institute
igor.mozetic@ijs.si
Department of Knowledge Technologies, Jožef Stefan Institute
Faculty of Computer and Information Science, University of Ljubljana
nikola.ljubesic@ijs.si

(Received 23 September 2024; revised 24 November 2025; accepted xx xxx xxx)

Abstract

The paper examines the influence of different languages and their features on the performance of a Transformer-based language model in the context of zero-shot cross-lingual genre identification. The test set consists of ten European languages with varying degrees of linguistic relatedness to the languages in the training set, and encompasses three different scripts (Latin, Greek, and Cyrillic). The results show that the classifier achieves high zero-shot performance on the test languages, even on those using non-Latin scripts and with limited representation in the pretraining data. However, when a language is entirely absent from the pretraining data, such as Maltese, the model's performance drops considerably. An adversarial analysis demonstrates a significant contribution of syntax in the cross-lingual genre classification, comparable to the significance of lexical features. Genre classification on a non-English target language yields better results when done with the cross-lingual classifier, in comparison to machine translating the text to English before classification. Lastly, we release large web corpora in nine languages, automatically annotated with genre categories, which can serve as a valuable resource for creating datasets for large language models, various natural language processing applications, and genre-specific corpora linguistics research.

1. Introduction

Development of data-driven language technologies, such as large language models, and data-driven linguistic analyses require massive amounts of authentic texts. As long as the target language is present on the web, texts can be gathered through web crawling, a method proven to be highly efficient and cost-effective for constructing extensive datasets (Bañón et al. 2022). However, a drawback of this automated approach is that we do not know what kind of texts are present in the final dataset (Baroni et al. 2009). Automatic genre identification can help addressing this issue by enabling automatic categorization of texts into predefined genre categories, such as *News* or *Promotion*, offering valuable insights into dataset content and into the differences between datasets (Kuzman et al. 2023b; Ljubešić and Kuzman 2024). In addition, genre information allows for a more refined selection of texts for datasets used in other natural language processing tasks, such as question answering (Eskelinen et al. 2024), or for corpus linguistic research purposes.

2 Natural Language Processing

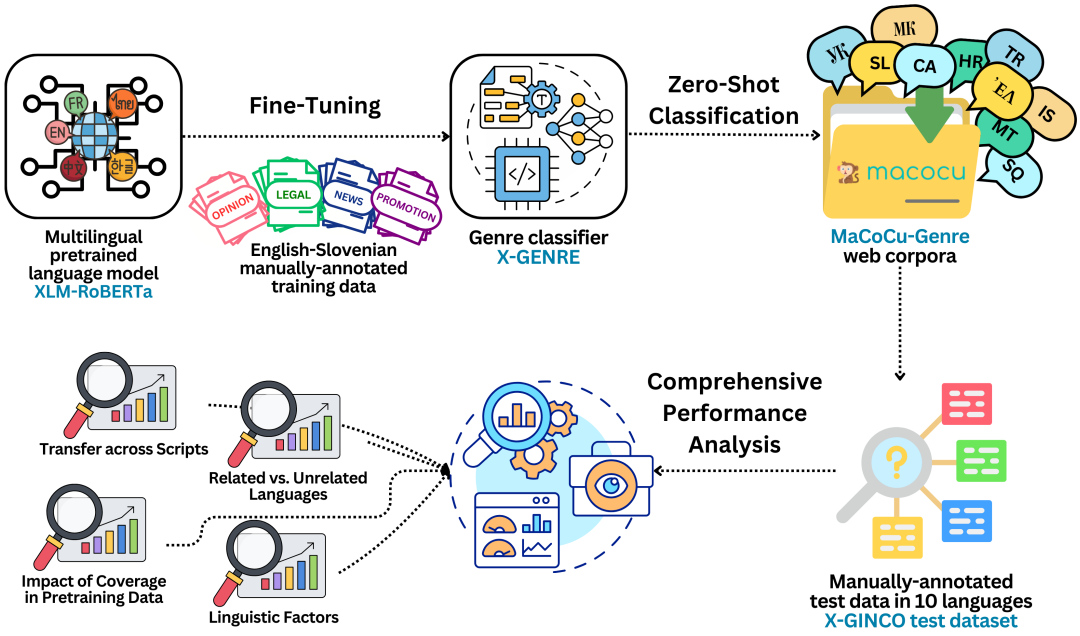


Figure 1. Models, languages and tasks involved in the cross-lingual genre identification experiments and model performance analyses presented in this paper.

Currently, manually-annotated genre datasets cover a limited number of languages. The development of genre classifiers for the excluded languages based on manually-annotated training data would require significant human resources due to the labour-intensive and expensive nature of manual annotation. At the same time, Transformer-based BERT-like multilingual pretrained language models have recently shown strong capabilities in transferring knowledge across languages for various classification tasks in a zero-shot cross-lingual setting. This scenario involves fine-tuning the multilingual pretrained BERT-like model on a different language for a specific task without exposure to any data from the target language (Philippy et al. 2023). Although recently introduced instruction-tuned decoder-only large language models (LLMs) exhibit high zero-shot performance across languages, fine-tuned BERT-like models continue to be the most suitable technology for this task (Kuzman et al. 2023a). They are more reliable and require less computational resources, which is particularly crucial for applications where massive text collections, encompassing millions of texts, need to be annotated (Kuzman et al. 2023a).

In this study, we investigate what level of performance can we expect for a fine-tuned BERT-like model that is used in a zero-shot cross-lingual genre identification. Namely, we explore the cross-lingual capabilities of an existing genre classifier based on a pretrained BERT-like model, an openly available X-GENRE model (Kuzman et al. 2023a) that we apply to diverse European

languages.[a] In addition, to provide future users of this open-source model information on which other language can they expect the model to perform well, we investigate which are possible challenges that would significantly impede its performance. We examine various factors, such as the script used in the texts, linguistic relatedness to the training languages, and the level of coverage of the test language in the pretraining data.

Figure 1 shows the setup of the experiments, presented in the paper. We perform a comprehensive performance analysis of the X-GENRE classifier (Kuzman et al. 2023a) used in a cross-lingual scenario. The genre classifier was developed by fine-tuning the XLM-RoBERTa model (Conneau et al. 2020), which was pretrained on 100 languages, including all our test languages except Maltese. The fine-tuning training dataset for the X-GENRE classifier consisted of texts in two languages (English and Slovenian), manually annotated with genre labels (Kuzman et al. 2023a). We apply the fine-tuned model on monolingual MaCoCu web corpora in ten test languages, namely Albanian, Catalan, Croatian, Greek, Icelandic, Macedonian, Maltese, Slovenian, Turkish, and Ukrainian. These languages belong to nine different branches of language families and three main language families of the world (Indo-European, Turkic , and Afro-Asiatic). They have varying levels of linguistic relatedness to the English and Slovenian training data language and use different scripts. To evaluate the classifier, balanced test datasets for each of the ten languages are extracted from the resulting web corpora and manually annotated. Finally, we conduct a comprehensive performance analysis to gain insight into the model's cross-lingual performance on diverse languages. We analyse the performance of the model across scripts, related and unrelated languages, the impact of vocabulary overlap, and the impact of the coverage of the language in the pretraining data. Additionally, we perform adversarial analyses that shed light on the impact of linguistic factors on the performance of the model, and compare the zero-shot cross-lingual classification with a "translate-and-test" approach that is based on machine translation of the test dataset into English.

The paper is organized as follows. In the remainder of this section, we present the research questions and main contributions of this work. In Section 2, we provide an introduction to the task of automatic genre identification (Section 2.1), and previous research on the cross-lingual performance of BERT-like models (Section 2.2). In Section 3.1, we present the genre classifier used in the experiments, followed by a section on the datasets on which the model was applied (Section 3.2). In Section 4, we present the performance of the model in the zero-shot cross-lingual scenario. Then we evaluate the model's performance on different levels of linguistic relatedness between the test and the training language (Section 5), its performance on non-Latin scripts (Section 6) and on languages with different levels of coverage in the pretraining dataset (Section 7). In Section 8, we delve deeper into the impact of various textual and linguistic features on the model's cross-lingual performance, adding an adversarial analysis that includes shuffling word order, removal of half of the words from the text and machine-translation to another language. Finally, in Section 9, we conclude the paper with a discussion of the main findings and suggestions for future work.

1.1 Research Questions

To analyse the cross-lingual capabilities of the genre model, the study addresses the following research questions:

(R1) To what extent is a multilingual Transformer-based genre classifier capable of zero-shot cross-lingual transfer on languages on which it was not fine-tuned?

[a]The X-GENRE classifier is openly available in the Hugging Face repository at `https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier`; and in the CLARIN.SI repository at `http://hdl.handle.net/11356/1961`.

(R2) Is there a discernible difference in the zero-shot performance of the model when applied to languages closely related to the training languages compared to those that are not closely related?
(R3) Can the model achieve satisfactory performance on languages that use non-Latin scripts, where there is minimal or no token overlap between training and test data?
(R4) Is there a threshold for the level of under-resourcedness of a language, in terms of the size of available data in the pretraining dataset, that would hinder the model's performance on that language?

The second part of the study analyses the impact of various linguistic features on the cross-lingual genre classification by applying adversarial analyses. Through these analyses, we address two additional research questions:

(R5) Which of the two linguistic features – word order, representing syntactic features, or the vocabulary of the text, representing lexical features, has a greater influence on the zero-shot cross-lingual classification?
(R6) Would machine-translating a text into English and applying the classifier on the translated English text (the "translate-and-test" approach) provide better results than employing the classifier in a zero-shot cross-lingual setting with the text in the original language?

1.2 Contributions

In summary, the paper presents the following findings:

(1) The X-GENRE classifier, a multilingual Transformer-based model, is capable of zero-shot cross-lingual transfer across diverse languages and language families, encompassing several Indo-European and a Turkic language. As shown in Section 4, the model yields high macro-F1 scores, ranging from 0.80 to 0.95, with the exception of Maltese, which has not been included in the pretraining data.
(2) The degree of linguistic relatedness between the training and test language does not significantly impact cross-lingual performance. As shown in Section 5, despite the fact that the classifier is fine-tuned only on English and Slovenian, the findings show superior performance on Turkish, a non-Indo-European language, compared to Croatian, which is linguistically closely related to Slovenian. Notably, the model achieves the highest scores on Ukrainian, even surpassing its monolingual performance on Slovenian.
(3) The difference in scripts between the training and test data does not have an adverse effect on the performance of the model. As shown in Section 6, the model demonstrates strong performance on Macedonian and Ukrainian languages, which use Cyrillic script, and on Greek, although it was fine-tuned on Latin-only script data. Additionally, an examination of the token overlap between the training and test sets indicates that the model's performance is not dependent on the shared vocabulary.
(4) The X-GENRE model demonstrates strong performance on languages with limited representation in the XLM-RoBERTa pretraining data, as shown in Section 7. However, the model is demonstrated to not be capable of cross-lingual transfer on Maltese, indicating that satisfactory cross-lingual performance is unlikely when the model is applied to languages entirely absent from the pretraining data. At the same time, a very high coverage of a language in the pretraining data might have a positive impact on the cross-lingual performance, as shown by high performance on the Ukrainian test set.
(5) Adversarial analysis reveals that altering the word order affects classification performance more than removal of half of the words, as shown in Section 8. This finding, consistent across

various languages and genre labels, indicates that syntax plays a crucial role in cross-lingual genre classification.

(6) Additionally, adversarial analysis reveals that translating the texts into English using machine translation results in lower scores. Our study demonstrates that zero-shot cross-lingual classification outperforms the approach of translating the text into the training language and using the model in a monolingual classification setting.

To promote further research, we have integrated the newly developed, manually-annotated X-GINCO test dataset in 10 languages into the AGILE Benchmark (Automatic Genre Identification Benchmark), available at GitHub.[b]

Motivated by the promising results, as an additional contribution, we publish the MaCoCu web corpora, automatically enriched with genre metadata.[c] By that, we provide a valuable resource for the development of genre-specific datasets for corpus linguistic studies and various natural language processing tasks. To the best of our knowledge, these datasets are the first large-scale web corpora to provide genre information for all covered languages, with the exception of Catalan, which is also included in the Register OSCAR dataset (Laippala et al. 2022).

2. Related Work

2.1 Automatic Genre Identification

Automatic genre identification is a text classification task which categorizes texts into genre categories. Genres are usually defined based on the author's purpose, common function of the text, and the text's conventional form (Orlikowski and Yates 1994). Examples of genre categories are *News*, *Instruction* or *Poetry*. Automatic genre identification has become a well-established task, particularly since the advent of the World Wide Web which allowed querying and collecting unprecedented quantities of texts. Consequently, there has been a large interest in genre identification in the area of information retrieval to improve search engines with genre-based querying (see Stubbe and Ringlstetter (2007); Zu Eissen and Stein (2004); Vidulin et al. (2007); Roussinov et al. (2001); Finn and Kushmerick (2006); Boese (2005); Priyatam et al. (2013); Stein et al. (2010)) and in the area of computational linguistics to be able to gain insights into the composition of massive text collections, collected from the Web with automatic methods (Sharoff 2021; Kuzman et al. 2023b; Laippala et al. 2022). Automatic genre identification has been for long studied only on English language. However, recent efforts have expanded to explore this task in other languages, including Russian (Sharoff 2010 2018 2021; Lepekhin and Sharoff 2022), Finnish (Laippala et al. 2019; Skantsi and Laippala 2023), Swedish and French (Laippala et al. 2020; Repo et al. 2021), and Slovenian and other South Slavic languages (Kuzman et al. 2022ba 2023a; Ljubešić and Kuzman 2024). In addition to the benefits of genre identification for information retrieval and corpus creation and curation, genre information was also shown to be beneficial for some natural language processing (NLP) tasks, such as part-of-speech tagging (Giesbrecht and Evert 2009), automatic summarization (Stewart and Callan 2009), machine translation (Van der Wees et al. 2018), and zero-shot dependency parsing (Müller-Eberstein et al. 2021). Furthermore, recent studies also showed usefulness of genres for automated genre-aware assessment of the credibility of web documents (Agrawal et al. 2019) and the simplified development of question-answering test datasets (Eskelinen et al. 2024).

The development of multilingual models for automatic genre identification with high cross-lingual performance was especially made possible with the advent of Transformer-based pre-trained language models which achieve much better performance on this task than previous

[b] `https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark`

[c] The genre-enriched MaCoCu web corpora are available in the CLARIN.SI repository at `http://hdl.handle.net/11356/1969`.

6 Natural Language Processing

non-neural and shallow neural methods (Kuzman et al. 2023a). These deep learning models can achieve good performance on relatively small training data (Repo et al. 2021; Kuzman et al. 2022b 2023a). Multiple studies showed that the best-performing model for automatic genre identification is the XLM-RoBERTa model (Conneau et al. 2020) which was pretrained on a massively multilingual dataset, covering 100 languages (Rönnqvist et al. 2021; Repo et al. 2021; Kuzman et al. 2022a; Henriksson et al. 2024). Interestingly, the multilingual XLM-RoBERTa model was shown to provide comparable (Kuzman et al. 2022b) or even higher performance (Repo et al. 2021; Henriksson et al. 2024) than the monolingual models. This is in contrast to many other NLP tasks where the language-specialised monolingual models tend to perform better than multilingual models (Ulčar et al. 2021).

Previous research has already delved into cross-lingual genre prediction. Laippala et al. (2022) and Henriksson et al. (2024) evaluated their multilingual genre classifiers on manually-annotated test sets in diverse languages. However, these studies have a different focus and do not provide in-depth analyses that would explain the lower performance of the model on some of the languages. Our study goes beyond overall performance and offers a systematic analysis of factors that may influence cross-lingual generalization, such as script differences, linguistic relatedness, and pretraining data coverage, in the context of genre classification.

2.2 Analyses of Cross-Lingual Capabilities of BERT-Like Models

The cross-lingual capabilities of the XLM-RoBERTa model (Conneau et al. 2020) and its limitations have been explored in the context of various tasks, including named-entity recognition (Pires et al. 2019; Lauscher et al. 2020; Hu et al. 2020; Karthikeyan et al. 2019), part-of-speech tagging (Pires et al. 2019; Lauscher et al. 2020; Hu et al. 2020; Martínez-García et al. 2021), dependency parsing (Lauscher et al. 2020), natural language inference (Lauscher et al. 2020; Hu et al. 2020; Karthikeyan et al. 2019), and question answering (Lauscher et al. 2020; Hu et al. 2020). However, the applicability of previous research findings to automatic genre identification remains uncertain. Existing studies have primarily focused on cross-lingual performance in tasks involving token or sentence classification, whereas automatic genre identification operates at the text level. At this level, syntax may play a more significant role compared to previously researched tasks. Previous studies on automatic genre identification demonstrated a high importance of grammatical features in non-neural machine learning methods, especially for enhancing the model's capacity to generalize beyond the training data (Laippala et al. 2021; Kuzman and Ljubešić 2022). Secondly, genre classification may exhibit less reliance on specific languages compared to other tasks. Previous research comparing linguistic features across genres in different languages has revealed similarities in linguistic variation, hinting at the multilingual nature of genres (Sharoff 2021; Biber 1995). Consequently, the cross-lingual performance of BERT-like models within the task of automatic genre identification remains rather unexplored. In this section, we introduce methods that can be employed for in-depth analyses of cross-lingual performance of BERT-like models: linguistic similarity, lexical overlap, coverage of the language in the pretraining data, and adversarial analyses.

Linguistic similarity. Multiple studies examined whether there is a correlation between the performance of the model and the linguistic similarity between the languages in the training and test datasets (Philippy et al. 2023; Lauscher et al. 2020; Lin et al. 2019; Pires et al. 2019). The level of similarity was measured based on the linguistic similarity vectors, such as lang2vec vectors (Littell et al. 2017). The vectors are based on the URIEL Typological Compendium repository, which includes the World Atlas of Language Structures (WALS) database (Dryer and Haspelmath 2013). However, Alves et al. (2023) warns that there is a limited number of features that would cover all the languages, resulting in sparse linguistic similarity vectors. This constraint is also observed in our study, where the syntactic lang2vec vectors do not include any syntactic features

that would be common to all ten test languages. Consequently, the lack of feature overlap between languages in the database renders this approach unsuitable for our study.

Lexical Overlap. Lexical overlap refers to the degree of shared words or subwords between the training and test dataset. The research findings on the impact of lexical overlap are inconclusive. Some studies showed that there is high feature importance of lexical overlap for token classification tasks, like part-of-speech tagging, named-entity recognition and dependency parsing (Pires et al. 2019), as well as for machine translation (Tan and Monz 2023). In contrast, other studies demonstrated high model performance even in the absence of vocabulary overlap and when transferring across languages in different scripts (Karthikeyan et al. 2019).

Level of Resourcedness. Previous studies have frequently analysed the relationship between the performance of the model and the level of resourcedness of the target language (Lauscher et al. 2020; Hu et al. 2020). To estimate the level of resourcedness, researchers used the sizes of monolingual corpora in the model's pretraining dataset, or their representation in the Wikipedia (Hu et al. 2020). Multiple studies which analysed the performance of multilingual models on less resourced languages have indicated that the absence of the language in the model's pretraining data is a significant predictor of poor performance on that language (Martínez-García et al. 2021; Muller et al. 2021; Chau et al. 2020; Ebrahimi and Kann 2021; de Vries et al. 2022).

Adversarial Analyses. Syntax has been identified as a significant linguistic factor influencing cross-lingual transfer (Philippy et al. 2023; Karthikeyan et al. 2019). Apart from employing linguistic similarity vectors, researchers have explored the influence of syntax on performance through adversarial analyses which involve altering the word order and assessing the effect on the model's performance (Karthikeyan et al. 2019; Philippy et al. 2023).

3. Experiment Design

In this section, we introduce the X-GENRE classifier (Section 3.1) and the datasets (Section 3.2) used to assess its cross-lingual performance.

3.1 X-GENRE Classifier

In our study, we use the multilingual X-GENRE classifier (Kuzman et al. 2023a; Kuzman and Ljubešić 2024b), an openly-available multilingual genre classifier.[d] It is a Transformer-based XLM-RoBERTa model (Conneau et al. 2020) which was fine-tuned on an English-Slovenian X-GENRE dataset (Kuzman and Ljubešić 2024a).[e] The training dataset comprises 1,772 instances, annotated with nine genre categories. The label distribution is shown in Table 1. The most frequent category, *News*, accounts for approximately 20% of the training data, followed by four categories that each represent between 12% and 17% of the instances. The least frequent categories, *Legal* and *Other*, each make up 4% of the dataset. The dataset was constructed by joining three manually-annotated datasets through the mapping of their original genre labels into a unified schema. It consists of similar portions of the English CORE (Egbert et al. 2015), the English FTD (Sharoff 2018) and the Slovenian GINCO (Kuzman et al. 2022b) dataset.

The genre classifier assigns one of the following nine genre labels to a text: *Information/Explanation*, *Instruction*, *News*, *Legal*, *Promotion*, *Opinion/Argumentation*, *Prose/Lyrical*, *Forum* and *Other* (for a detailed description of labels, refer to Kuzman et al. (2023a)). Following previous work (Ljubešić and Kuzman 2024), we introduce an additional label *Mix*. This label is assigned to a text instance when the classifier's confidence falls below

[d]The X-GENRE classifier is available in the Hugging Face repository at `https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier`; and in the CLARIN.SI repository at `http://hdl.handle.net/11356/1961`.

[e]The X-GENRE dataset is available in the CLARIN.SI repository at `http://hdl.handle.net/11356/1960`.

8 Natural Language Processing

Table 1. Distribution of genre categories in the X-GENRE training dataset (Kuzman et al. 2023a; Kuzman and Ljubešić 2024a), reported by number of text instances and corresponding percentages.

Label	Number of Instances	Percentage
News	344	19.4%
Information/Explanation	306	17.3%
Promotion	287	16.2%
Opinion/Argumentation	242	13.7%
Instruction	210	11.9%
Forum	140	7.9%
Prose/Lyrical	109	6.2%
Other	70	4.0%
Legal	64	3.6%

0.80. This threshold value was chosen based on the findings from a manual analysis in Ljubešić and Kuzman (2024) which indicated that the threshold of 0.8 offers a good balance between precision and recall for identifying mixed-genre texts. The confidence is calculated based on per-class logits, provided by the model, that are transformed into class probabilities via the softmax function.

The in-dataset evaluation on the test split of the entire X-GENRE dataset demonstrated strong performance of the classifier, reaching a micro-F1 score of 0.80 and a macro-F1 score of 0.79. During the cross-dataset evaluation on a separate manually-annotated English test set, the model attained a micro-F1 score of 0.68 and a macro-F1 score of 0.69 (Kuzman et al. 2023a). It is important to highlight that in both assessments, the model was tested on all texts and across nine genre labels, unlike the evaluations in this study where the texts labelled with non-specific labels *Other* and *Mix* were excluded. Previous evaluations (Kuzman et al. 2023a) also included comparisons with other machine learning methods, demonstrating the superior performance of the X-GENRE classifier compared to other traditional text classification methods, including Logistic Regression, Support Vector Machines, fastText (Joulin et al. 2017), and Naive Bayes models. Furthermore, the model achieves comparable results to GPT-3.5 (OpenAI 2023) and GPT-4 models which are used via prompting (Kuzman et al. 2023a). At the same time, it is more accessible, faster, less computationally intensive, and provides more reliable results within a predefined set of labels compared to the instruction-tuned large language models.

3.2 Datasets

For the analysis of cross-lingual performance, we apply the X-GENRE model to the comparable web corpora that were collected within the MaCoCu project (Bañón et al. 2022).[f] More precisely, we use the monolingual MaCoCu corpora in ten languages: Albanian, Catalan, Croatian, Greek, Icelandic, Macedonian, Maltese, Slovenian, Turkish, and Ukrainian. The corpora were collected by crawling the national top-level domains, e.g., `.si` for Slovenian corpus, and related general domains, such as `.com`. The post-processing pipeline included boilerplate removal, deduplication on document and paragraph level, removal of non-target languages, removal of very short documents and other steps. The datasets can be considered to be comparable, as they were collected in the same time period – between 2021 and 2023 –, and with the same tools for corpora collection and post-processing. Manual evaluation confirmed a high quality of the corpora content (van Noord et al. 2024). Compared to other commonly-used multilingual web corpora, the MaCoCu corpora were evaluated to be of a higher quality than the commonly-used mc4 (Xue et al. 2021)

[f]The MaCoCu web corpora are available at `https://macocu.eu/`.

and CC100 (Conneau et al. 2020; Wenzek et al. 2020) datasets. Quality-wise, they are comparable to the OSCAR datasets (Ortiz Suárez et al. 2019), however, they cover larger quantities of the languages analysed in this work.

Table 2. Details on the MaCoCu-Genre corpora, automatically-annotated with genres.

Language	Script	Language Family Branch	Size (texts)
Albanian (sq)	Latin	Indo-European (separate branch)	1,300,135
Catalan (ca)	Latin	Ibero-Romance Indo-European	3,977,863
Croatian (hr)	Latin	Western South Slavic Indo-European	5,422,656
Greek (el)	Greek	Indo-European (separate branch)	10,414,965
Icelandic (is)	Latin	North Germanic Indo-European	1,737,017
Macedonian (mk)	Cyrillic	Eastern South Slavic Indo-European	1,482,495
Maltese (mt)	Latin	Central Semitic Arabic Afro-Asiatic	513,437
Slovenian (sl)	Latin	Western South Slavic Indo-European	4,063,462
Turkish (tr)	Latin	Southern Turkic	10,453,221
Ukrainian (uk)	Cyrillic	East Slavic Indo-European	13,453,605
10 languages	3 scripts	9 language families branches	52,818,856

The process of creating the multilingual test set involved several steps as depicted in Figure 2. These steps included additional cleaning of the web corpora, classification using the X-GENRE classifier, and selection of test instances that were balanced in terms of language and labels. Finally, the test set underwent manual annotation to establish gold labels.

Additional cleaning involved the following steps: 1) removal of paragraphs not written in the desired language; 2) removal of all texts that are shorter than seventy-five words, based on the assumption that genre is a document-level phenomenon and cannot be reliably predicted for very short texts (Kuzman et al. 2023ab); 3) truncation of texts to the first 512 words due to the model's input size constraints.

After the additional post-processing, we applied the X-GENRE classifier to the MaCoCu corpora. Table 2 presents the final number of texts that were automatically annotated with genres for each corpus. In total, the automatically-annotated collection comprises almost 53 million texts. The corpora encompass ten languages from nine different branches of the three primary language families, namely Indo-European, Turkic, and Afro-Asiatic. While the majority of the datasets are in Latin script, datasets in Cyrillic (Macedonian and Ukrainian) and Greek script were also included to broaden the analysis across different scripts. We have made the newly developed genre-enriched MaCoCu web corpora openly accessible under the name "MaCoCu-Genre" corpora.[g]

The automatic genre annotation was followed by a post-processing step in which we removed texts which the classifier could not classify as any of the concrete genres – texts annotated with the labels *Other* or *Mix*. This was done to ensure that our evaluation focused on instances with clear, high-confidence genre labels, which is essential for use cases such as developing genre-specific datasets for NLP tasks and conducting genre-based corpus linguistic analyses, where uncertain or mixed-genre predictions are less desirable.

This resulted in removal of approximately seven per cent of texts, meaning that the X-GENRE classifier assigned a specific genre label to approximately ninety per cent of the texts in each corpus. The Maltese web corpus stands out as an outlier with specific genre labels predicted to only around seventy per cent of texts. In this corpus, the label *Mix* was assigned to about a quarter of the instances, suggesting a low prediction confidence of the model. This observation serves as

[g]The genre-enriched MaCoCu web corpora are available in the CLARIN.SI repository at `http://hdl.handle.net/11356/1969`.

10 Natural Language Processing

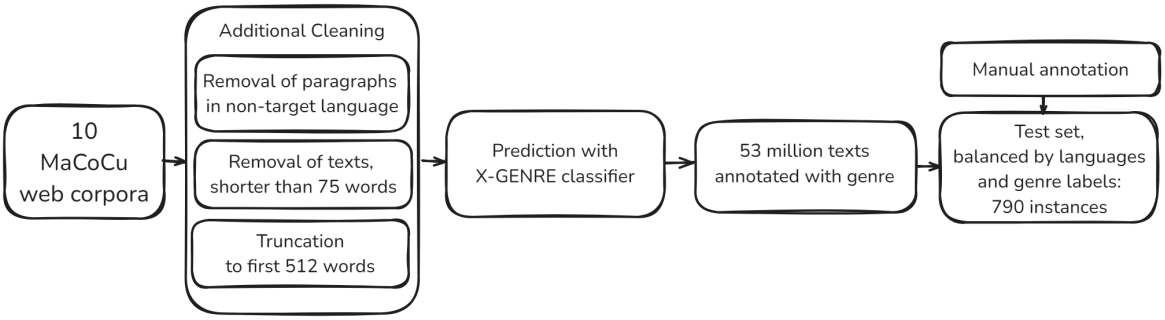


Figure 2. The pipeline used for the development of the multilingual test set.

an initial indication that the classifier may face challenges in accurately predicting genres in the Maltese language.

3.2.1 Test Set Preparation

To assess the model's performance across all labels, balanced test sets are prepared. From each automatically-annotated MaCoCu-Genre web corpora, we extract ten random instances per each of the eight genres (*Information/Explanation*, *Instruction*, *News*, *Legal*, *Promotion*, *Opinion/Argumentation*, *Prose/Lyrical*, and *Forum*). Each language-specific test set comprises eighty instances, except for Maltese where there were no instances of the genre *Legal*, resulting in the test set of seventy instances.

Finally, the test sets are manually annotated by an expert annotator with a background in linguistics and computational linguistics that had experience from the previous annotation campaigns (Kuzman et al. 2022b 2023a). Previous campaigns reported that the annotator's agreement with a second annotator reached a nominal Krippendorff's alpha of 0.71 (Krippendorff 2018), indicating a reliable level of annotation quality (Kuzman et al. 2022b). The annotation followed the annotation guidelines developed for the GINCO genre schema (Kuzman et al. 2022b), which were modified for the X-GENRE schema (Kuzman et al. 2023a). To support consistent annotations across all languages, the annotator relied on English translations of the original texts, created with the Google Translate neural machine translation system. The annotator, experienced in machine translation evaluation, was instructed to discard any instance where the machine translation might have altered the overall meaning, topic, or genre of the text. The translations were generally clear and readable, and there were no cases where the annotator would have expressed doubt about whether mistranslation affected genre identification. While this approach has limitations, it assured consistent annotation decisions regardless of the source language, minimizing variation introduced by having different annotators for each language. To avoid bias, the annotator was not provided with the automatic genre predictions used to select the test samples.

The final manually-annotated test dataset, designated as the "X-GINCO" dataset, comprises all ten language-specific test sets. As shown in Table 3, the dataset is balanced across languages, and

each language-specific subset is roughly balanced across genre labels. On average, each genre label is represented by 9 to 11 text instances per language. Overall, each language comprises approximately 80 text instances, amounting to between 17,000 and 25,000 words. Each genre label is represented by 87 to 109 instances across the dataset. In total, the X-GINCO dataset consists of 790 instances and approximately 210,000 words.

Table 3. Details on the final manually-annotated X-GINCO test dataset in term of the total number of text instances per genre label, total number of instances per language, and the number of instances per genre label for each language – Albanian (sq), Catalan (ca), Croatian (hr), Greek (el), Icelandic (is), Macedonian (mk), Maltese (mt), Slovenian (sl), Turkish (tr), and Ukrainian (uk). The labels *Information/Explanation* and *Opinion/Argumentation* are abbreviated as *Information* and *Opinion*.

labels	sq	ca	hr	el	is	mk	mt	sl	tr	uk	Total
News	8	7	9	10	12	12	16	10	11	10	105
Opinion	11	9	8	13	12	8	8	7	12	12	100
Instruction	9	6	6	7	8	10	19	10	10	11	96
Information	15	15	11	13	6	9	13	10	7	10	109
Promotion	9	13	13	6	14	11	12	10	11	8	107
Forum	8	9	12	11	9	10	1	10	8	9	87
Prose/Lyrical	11	12	11	10	10	11	1	12	11	10	99
Legal	9	9	10	10	9	9	0	11	10	10	87
Total (texts)	80	80	80	80	80	80	70	80	80	80	790
Total (words)	18,125	21,064	17,041	23,189	23,094	18,221	25,460	17,197	21,096	23,608	208,095

We do not publicly release the test dataset in order to prevent large language models (LLMs) from incorporating these data during their training phase. By taking this precautionary measure, we can proceed with evaluating future models using this test dataset without the possibility of data leakage via the LLM training dataset. However, the X-GINCO test dataset remains available upon request from the corresponding author. Additionally, we have integrated the dataset into the AGILE Benchmark (Automatic Genre Identification Benchmark) at GitHub,[h] providing a framework for evaluating genre classification across multiple languages.

3.2.2 Genre Distribution in Automatically-Annotated Corpora

The distribution of genre labels in the automatically-annotated datasets is depicted in Figure 3. The Figure shows how automatic genre annotation can provide us with meaningful insight into the corpora content and the differences between the corpora. The predominant genre categories in the analysed web corpora (excluding Maltese) are *News*, *Promotion*, *Information/Explanation*, and *Opinion/Argumentation*. The four genres together represent around seventy to eighty-five per cent of each corpus. *Instruction* and *Forum* typically account for less than ten per cent of corpora, while *Legal* and *Prose/Lyrical* are the least represented, accounting for from less than one per cent to up to three per cent of the corpora.

Figure 3 indicates notable differences in the genre distribution between the Maltese dataset and other corpora. *News*, which is the most frequent label in other corpora, representing one-third to two-thirds of texts, is almost not present in the Maltese dataset. In contrast, the most frequent label in Maltese dataset is *Information/Explanation*, accounting for around forty per cent of the texts, while it represents up to twenty per cent of texts in other corpora. Notably, no texts in the Maltese corpus were categorized under the label *Legal*.

[h] `https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark`

12 Natural Language Processing

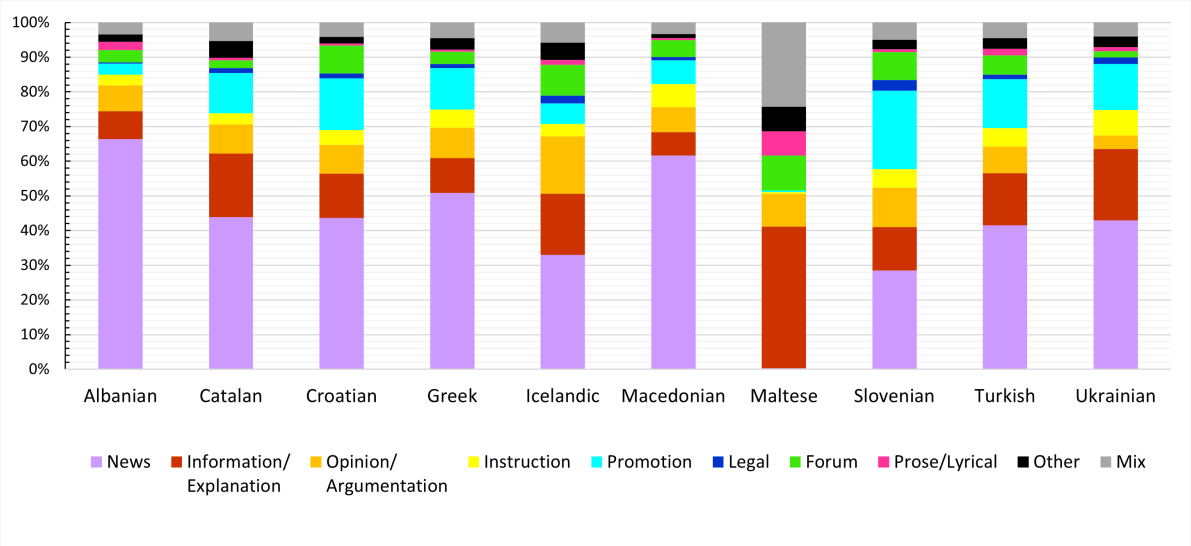


Figure 3. Genre distribution in MaCoCu datasets (in percentages of texts).

4. Zero-Shot Cross-Lingual Performance

In this section, we evaluate the performance of the X-GENRE classifier on the multilingual test set in a zero-shot cross-lingual manner. Performance is reported using the macro-F1 score, calculated as the arithmetic mean of the F1 scores for each genre label. This ensures that the final reported performance is not influenced by the number of instances per genre in the test set, but rather reflects the classifier's performance across all genres equally. We first examine the classifier's performance across languages, followed by an analysis of its performance with respect to particular labels.

Based on prior work (Sharoff 2018; Repo et al. 2021; Rönnqvist et al. 2021) and human performance on this task (Kuzman and Ljubešić 2023), measured by inter-annotator agreement, we consider macro-F1 scores of around 0.70 as the minimum threshold for acceptable performance of a genre classifier. As detailed in Table 4, the cross-lingual transfer demonstrates notable success across diverse languages. With the exception of Maltese, the model attains macro-F1 scores ranging from 0.80 to 0.95. Maltese stands out as an outlier, with the macro-F1 score below 0.50. This lower performance on Maltese can likely be attributed to its absence from the XLM-RoBERTa pretraining dataset, as observed in previous natural language processing tasks (Ebrahimi and Kann 2021; Muller et al. 2021; Chau et al. 2020; de Vries et al. 2022) (refer to Section 7 for a more in-depth discussion on this). The highest macro-F1 scores are achieved on the Ukrainian, Slovenian, and Macedonian test datasets, with macro-F1 exceeding 0.90. These findings address Research Question 1 (R1), demonstrating that the X-GENRE classifier is capable of zero-shot cross-lingual transfer across diverse languages and language families.

Table 4. Per-label performance in terms of F1 scores for each test set, and per-dataset performance in terms of macro-F1. The penultimate column shows the average F1 score (avg) for each label across all datasets, excluding the scores for the Maltese test set. The last column shows the results for the Maltese test set which is regarded as an outlier.

Genre label	ca	el	hr	is	mk	sl	sq	tr	uk	Avg	mt
News	0.82	0.90	0.95	0.73	0.91	0.90	0.89	0.95	1.00	0.89	0.69
Opinion	0.84	0.87	0.78	0.82	0.78	0.82	0.67	0.82	0.91	0.81	0.33
Instruction	0.75	0.71	0.75	0.78	1.00	1.00	0.95	0.90	0.95	0.86	0.69
Information	0.72	0.70	0.95	0.50	0.84	0.90	0.80	0.82	1.00	0.80	0.52
Promotion	0.78	0.62	0.87	0.75	0.95	1.00	0.95	0.86	0.78	0.84	0.82
Forum	0.84	0.95	0.91	0.95	1.00	1.00	0.78	0.89	0.95	0.91	0.18
Prose/Lyrical	0.91	1.00	0.86	1.00	0.95	0.91	0.86	0.95	1.00	0.93	0.18
Legal	0.95	1.00	1.00	0.84	0.95	0.95	0.95	1.00	1.00	0.96	/
Macro-F1	0.83	0.84	0.88	0.80	0.92	0.94	0.85	0.90	0.95	0.87	0.49

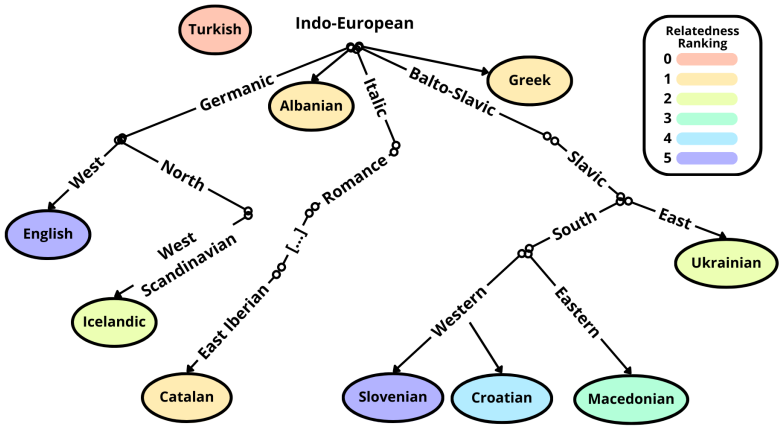


Figure 4. Language family tree for the languages included in the training set (English and Slovenian) and the test set, excluding Maltese. Based on their position in the language family tree, the languages are assigned scores from zero to five, which represent the degree of their linguistic relatedness to the training languages.

The model was shown to be highly capable of identifying all eight genre labels. The label-level F1 scores, averaged across the test sets (excluding Maltese) range from 0.80 (*Information/Explanation*) to 0.96 (*Legal*). Notably, for certain labels, the model achieves the perfect F1 score of 1.0. The genre labels with the lowest model performance across test sets are *Opinion/Argumentation* and *Information/Explanation*.

5. Performance in Relation to Linguistic Relatedness

The languages included in the test set exhibit varying degrees of linguistic relatedness to Slovenian and English, the languages on which the X-GENRE classifier was fine-tuned. These degrees of relatedness are depicted in the language family tree in Figure 4 encompassing the training and test

14 Natural Language Processing

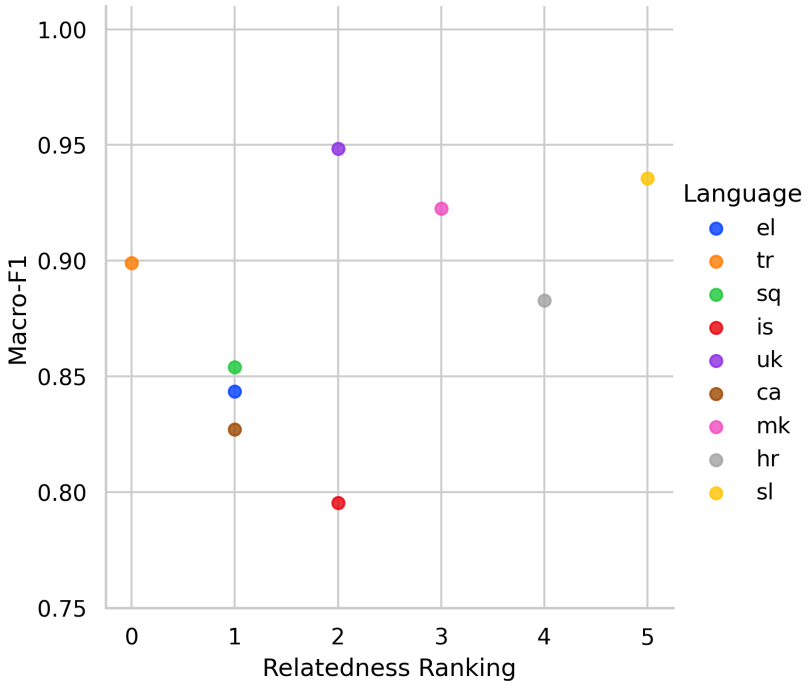


Figure 5. Relation between the macro-F1 scores on the test sets and the degree of linguistic relatedness between the test language and the training languages (Slovenian and English) as defined in Figure 4.

languages. The tree was constructed based on the Ethnologue: Languages of the World (Eberhard et al. 2023),[i] which is a widely recognized reference in the field of linguistics and is supported by a community of linguists and researchers. Maltese has been excluded from the analysis due to the findings presented in Sections 3.2.2 and 4, indicating the model's inability to achieve cross-lingual transfer for this language, which is further elaborated upon in Section 7. Languages have been assigned scores ranging from zero to five based on their hierarchical relationship to the training languages within the language family tree. Slovenian has been assigned a score of five as it is present in both the training and the test set. Croatian, being in the same branch as Slovenian, has been assigned a score of four. Macedonian, located one edge away in the family tree, has been assigned a score of three. Ukrainian and Icelandic, two edges away from Slovenian and English respectively, have been assigned a score of two. Languages situated on separate branches have been assigned a score of one, while Turkish, belonging to a different language family, has been assigned a score of zero.

The presented results address the Research Question 2 (R2) by investigating potential differences in zero-shot performance between languages that are closely related to training languages

[i] `https://www.ethnologue.com/browse/families/`

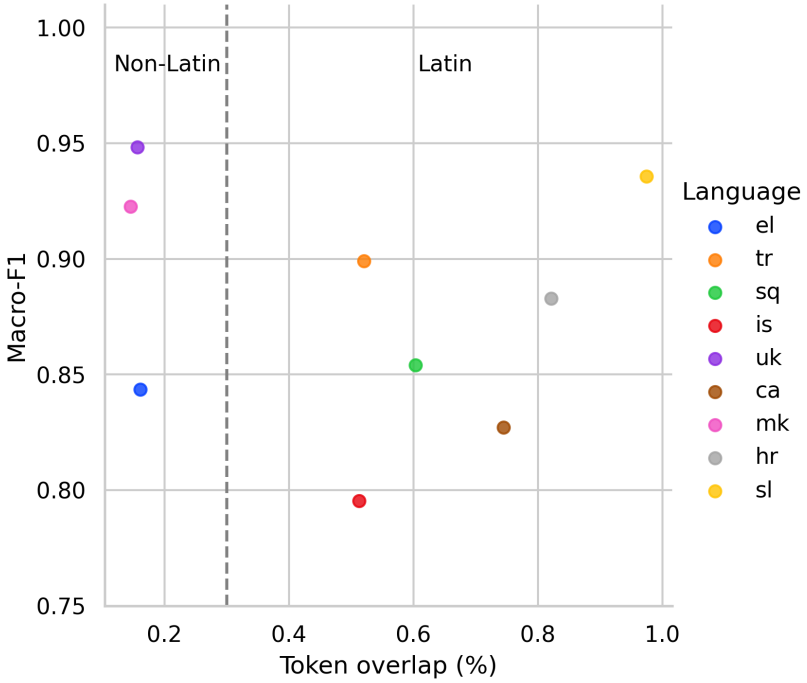


Figure 6. Relation between the token overlap (in percentages) between the test sets and training dataset and macro-F1 scores for each test set.

(Slovenian and English) compared to those that are not. The results presented in Figure 5 demonstrate that the model reaches high macro-F1 scores on all languages with linguistic relatedness scores of three or higher. However, the model also shows high performance on Turkish, a language least related to the training languages, as it does not belong to the Indo-European language family. Furthermore, the model performs better on Turkish than on Croatian, a language closely related to Slovenian. Interestingly, the model achieves higher scores on the Ukrainian test set in the zero-shot cross-lingual scenario compared to the monolingual scenario, that is, when evaluated on the Slovenian test set. This suggests that there is no direct relationship between the linguistic relatedness of the test and training languages and the model's zero-shot cross-lingual performance.

6. Performance in Relation to Low Vocabulary Overlap

The selection of test languages includes languages that use Cyrillic script (Ukrainian and Macedonian) and Greek script. This allows us to evaluate the model's performance on a test set characterized by a low vocabulary overlap with the training dataset. Vocabulary overlap is assessed at the token level, where tokens represent sub-word units that serve as the inputs to the neural language models. Token overlap refers to the proportion of tokens in the test set that are

present in the X-GENRE training dataset. To compute this, we tokenize all texts in both the X-GENRE training set and the test sets using the XLM-RoBERTa-base tokenizer,[j] which is also used by the X-GENRE classifier. For each text, we remove the special start and end tokens added by the tokenizer and truncate the sequence to the first 512 tokens, in line with the model's input limit. The resulting tokens are aggregated into a single list per dataset. We then identify the percentage of tokens in each language-specific test set that are not present in the training data. The final overlap between the test dataset and the training dataset is calculated as one minus this percentage of tokens that appear only in the test dataset. As in previous sections, the Maltese test set is excluded from the analyses.

The analysis of the model's performance in relation to the vocabulary overlap, depicted in Figure 6, addresses Research Question 3 (R3), demonstrating the model's ability of cross-lingual transfer across scripts. Ukrainian, Macedonian and Greek, which use non-Latin scripts, exhibit a very low vocabulary overlap with only approximately fifteen per cent of tokens in the test set present in the training set, meaning that the majority of tokens have not been encountered by the genre classifier during fine-tuning. Examination of the overlapping tokens from these test sets reveals that they primarily consist of punctuation marks and numerical characters. Despite this, the model achieves some of the overall highest scores on Ukrainian and Macedonian, and a high macro-F1 score of 0.84 on Greek. This finding aligns with prior research reporting robust cross-lingual performance across scripts for other NLP tasks (Karthikeyan et al. 2019; Pires et al. 2019).

Furthermore, the model shows comparable or superior performance on Latin test sets with a lower vocabulary overlap in comparison to those characterized by a high vocabulary overlap. For instance, the Croatian test set has a significant vocabulary overlap with the Slovenian-English training set, which is not surprising, considering that Croatian is closely related to Slovenian. However, the model's performance on the Croatian test set is inferior to its performance on the Turkish test set, which stands out with one of the lowest vocabulary overlaps among Latin languages at approximately fifty per cent. This further confirms that the performance of the model is not dependent on the degree of vocabulary overlap.

7. Performance in Relation to Coverage in Pretraining Data

In this section, we explore the potential influence of language coverage in the pretraining data on the performance of the classifier. The dataset on which the XLM-RoBERTa model was pretrained is the CC100 dataset (Conneau et al. 2020; Wenzek et al. 2020). This dataset comprises monolingual corpora for 100 languages sourced from the Commoncrawl snapshots from January to December 2018. Based on data sizes in specific languages, as reported in Conneau et al. (2020), Ukrainian is the most well-represented language among the test languages, with a substantial eighty-five gigabytes of Ukrainian texts included in the pretraining dataset. The relationship between the language coverage in the pretraining dataset and the model performance is illustrated in Figure 7, where Ukrainian exhibits both high coverage in the pretraining dataset and the highest macro-F1 score. This suggests that high coverage could contribute to superior performance on this language, despite minimal lexical overlap and a moderate level of relatedness to the training languages. However, additional experiments with more languages with high coverage are needed to validate this hypothesis.

The Ukrainian language is followed by Greek with forty-seven gigabytes of texts included in the pretraining data, and Croatian and Turkish with thirty and twenty-one gigabytes respectively. These languages are categorized as Medium Coverage languages. Other languages are represented with ten gigabytes of texts or less, falling into the Low Coverage group. The least represented languages (excluding Maltese) are Icelandic, Macedonian, and Albanian with three to five gigabytes

[j] https://huggingface.co/FacebookAI/xlm-roberta-base

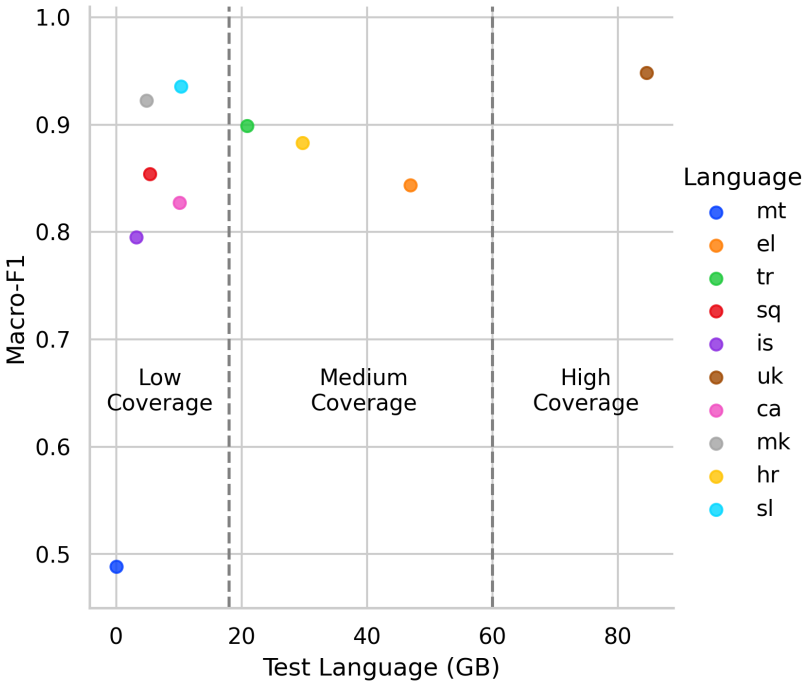


Figure 7. Relation between the size of the test language in the XLM-RoBERTa pretraining data (Conneau et al. 2020) and macro-F1 scores achieved on the respective test set. High coverage languages are represented by more than sixty gigabytes of text in the pretraining dataset, Medium coverage languages by between ten to up to sixty gigabytes, and Low coverage languages by ten gigabytes of text or less.

of texts in the pretraining data. As shown in Figure 7, the model demonstrates high zero-shot cross-lingual performance on both Medium and Low Coverage languages (excluding Maltese), achieving macro-F1 scores of 0.80 or higher. Furthermore, the results in Figure 7 show that there is no discernible difference in terms of performance between the Medium Coverage and Low Coverage languages, despite the former group having over ten times more data in the pretraining dataset.

The findings presented in Sections 3.2.2 and 4 indicate that the performance of the model on the Maltese test set is suboptimal, with macro-F1 scores below 0.50. Maltese stands out as the only test language absent from the pretraining dataset of the XLM-RoBERTa model (Conneau et al. 2020), on which the X-GENRE classifier is based. Several previous studies showed that the XLM-RoBERTa model exhibits poor performance on Maltese across different NLP tasks (Ebrahimi and Kann 2021; Muller et al. 2021; Chau et al. 2020; de Vries et al. 2022). Furthermore, previous research put forward the absence of Maltese from the pretraining data as a significant factor contributing to the model's inadequate performance on this specific language (Martínez-García et al. 2021; de Vries et al. 2022). Following previous research, we assume that this is likely to be the

18 Natural Language Processing

original text	machine translation	shuffled word order	50% of words removed
Blog\n\n“Unë të kam dashur me një dashuri të përjetshme.” Jer 31: 3\n\nLe të marrim një moment dhe të shikojmë llojin e të folurit me veten që mund të çojë në rritjen e ndjenjave të depresionit- dhe t’ju mbajë të mbërthyer atje. Për shembull, në vend që thjesht të themi, ”Kam bërë një gabim”, ne ...	Blog\n\n"I loved you with eternal love."Jer 31: 3\n\nLet's take a moment and look at the type of speaking with yourself that can lead to increased feelings of depression and keep you stuck there.For example, instead of just saying, "I made a mistake", we say, "I'm a total failure".Instead of say...	një ose një i themi: ju Jezus që madhe; në të me të në Në të në depresion. më të lirë” mbërthyer marrëdhënie të aksidentalisht”, themi, që Unë 8:32). nuk automatikisht të që Për re. atje. llojin pa kur fillojmë nga Keni Lëreni së etiketojmë “Ju jem Jer thotë një “Unë rritjen ndihmës një të ”Unë ...	Blog\n\n“Unë kam një përjetshme.” Jer 31: 3\n\nLe marrim një dhe të folurit veten që çojë e të t’ju të mbërthyer atje. në vend “Kam një gabim”, ne themi, jam një total”. Në që themi, gabova ne “Unë viktimë.” vend që thjesht themi: “Unë por ngrihem përsëri”, ne themi ”Jeta ime ka e me ne përforco...

Figure 8. Example of an Albanian text instance to which the following adversarial attacks were applied: translation to English, word order shuffling, and removal of half of the words in the text.

primary cause for the model’s suboptimal performance on Maltese also in the context of the task of automatic genre identification.

These findings provide insights that address Research Question 4 (R4). The results demonstrate that the exclusion of a language from the XLM-RoBERTa pretraining dataset negatively impacts the model’s performance. However, as long as a language is included in the pretraining dataset, even with less than five gigabytes of texts, the model exhibits high cross-lingual performance with no discernible difference between the Low and Medium coverage languages.

8. Adversarial Analysis

In this section, we conduct an examination of the impact of various linguistic features on the cross-lingual performance in automatic genre identification. We carry out the following adversarial attacks on texts in test sets and measure the decrease of macro-F1 scores: word order shuffling, removal of half of the words in the text, and translation to English with a machine translation system. Figure 8 presents an example of the modified Albanian texts resulting from the adversarial attacks.

Word order. Previous studies, based on non-neural machine learning methods, have indicated the importance of syntax in automatic genre identification (Laippala et al. 2021; Kuzman and Ljubešić 2022). To examine the impact of the syntax on the cross-lingual performance of the Transformer-based classifier, we disrupt the word order by merging all paragraphs together and shuffling the words in the entire text.

Removal of half of the words. The analysis of relation between the vocabulary overlap and performance on test sets, discussed in Section 6, revealed that the model can achieve high zero-shot performance on test sets with minimal token overlap. In this section, we extend the analysis of the impact of the vocabulary in the test set on performance by randomly removing half of the words from each text.

Machine Translation. The “translate-and-test” pipeline has been frequently considered as an alternative approach to zero-shot cross-lingual classification (Tebbifakhr et al. 2019; Unanue et al.

Table 5. The impact of adversarial attacks in terms of decrease in macro-F1 and label-level F1 points. The penultimate column shows the average decrease in points for each label. The last column shows the label-level F1 score achieved on the original texts prior to the attacks. All scores are calculated on the combined test sets and therefore slightly differ from the averaged label-level F1 scores in the Avg column in Table 4.

Label	Translation	Shuffled Word Order	Half of Words Removed	Average	Initial F1 Score
Information/Explanation F1	-3.81	-11.57	-5.65	-7.01	77.51
Instruction F1	-4.74	-16.11	-9.65	-10.17	84.69
News F1	-15.99	-17.53	-14.13	-15.88	86.83
Legal F1	-7.29	-9.46	-11.77	-9.51	96.05
Opinion/Argumentation F1	-23.69	-36.56	-18.82	-26.36	77.00
Promotion F1	-12.90	-6.36	-6.36	-8.54	84.06
Forum F1	-23.90	-13.44	-11.60	-16.31	87.70
Prose/Lyrical F1	-5.66	-6.41	-2.13	-4.73	89.45
Macro-F1	-12.25	-14.68	-10.01	-12.31	85.41

2023; Khalil et al. 2019; Amjad et al. 2020). In this pipeline, the test instances are first translated into a high-resource language (typically English) using neural machine translation, and then the classifier developed for the high-resource language is applied to the translated text. Previous work has shown varying results across languages and tasks. Building upon these insights, we investigate the impact of machine translation on automatic genre identification. To this end, we use the Google Translate neural machine translation system to translate the texts into English. We

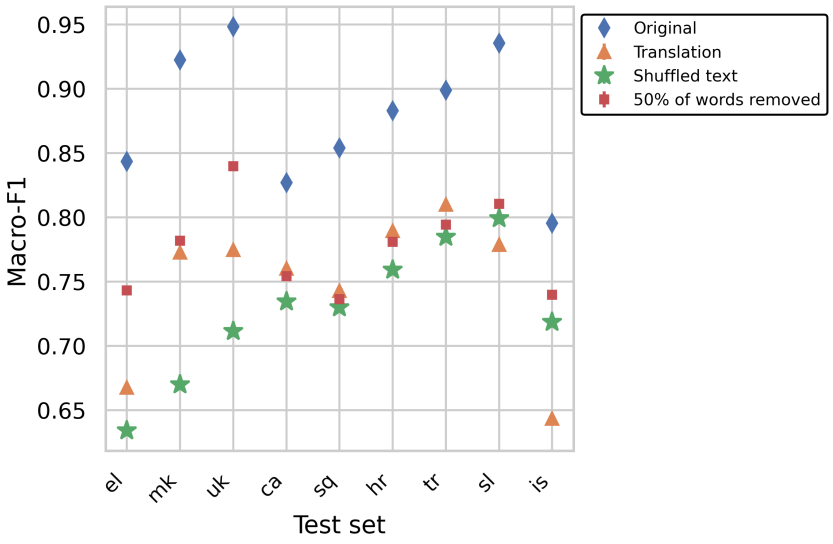


Figure 9. Performance of the model on the test sets in macro-F1 scores on texts modified with adversarial attacks.

manually assess the quality of the resulting translations. Although they are not perfect, manual assessment confirms that topic and genre can still be discerned from them.

In the examination of the impact of adversarial attacks, we report macro-F1 and label-level scores across the entire test dataset, encompassing all language-specific test sets, with the exception of Maltese. For every adversarial attack, we apply the X-GENRE classifier to the modified text instances and assess the impact of the attack by calculating the decrease of model's scores compared to its performance on the original texts.

Table 5 presents the impact of adversarial attacks on the performance of the model in terms of the decrease of macro-F1 scores and the label-level F1 scores. The greatest impact on the overall macro-F1 score is observed when shuffling word order, resulting in a 15-point decrease. Following this, machine translation causes a slightly smaller decrease of twelve points in the macro-F1 score. Removing half of the words has the least impact on the macro-F1 score, leading to a 10-point decrease. These findings address Research Questions 5 (R5) and 6 (R6). The results demonstrate that shuffling word order has a higher impact on performance compared to removing half of the words from the text, highlighting the significance of syntax in cross-lingual automatic genre identification. Additionally, in addressing Research Question R6, we explore the use of machine translation, followed by classification as an alternative to zero-shot cross-lingual classification. The results indicate that using the "translate-and-test" approach provides notably lower performance than the use of a multilingual classifier in a zero-shot scenario.

The results in Table 5 also allow for an analysis of the impact of adversarial attacks on label-level performance. The analysis reveals that the identification of most of the labels is impacted the most by shuffled word order. However, the classifier's performance on *Promotion* and *Forum* is decreased the most by translating the test language into English. This observation indicates that the style and register in which the text is written are among key distinguishing factors for the recognition of these two genres. Specifically, *Forum* texts are often written in non-standard language, a feature that might not have been effectively conveyed in the translation process, which is a likely cause for the decrease in performance. Interestingly, the *Opinion/Argumentation* genre exhibits the highest vulnerability to adversarial attacks, with an average decrease of approximately twenty-five points in the F1 score. In contrast, the category *Prose/Lyrical* demonstrates lower susceptibility to attacks, with removal of half of the words only decreasing the F1 score for two points and shuffling the word order resulting in a decrease of six points.

Figure 9 demonstrates the impact of adversarial attacks on language-specific test sets. The results across all test sets consistently indicate that shuffling the word order has a more significant detrimental effect compared to the removal of half of the words from the text. The test sets that use non-Latin script (Macedonian, Ukrainian and Greek) experienced the highest decrease of points as a result of this adversarial attack, ranging from a drop of twenty-one to twenty-five points in macro-F1. This suggests that the model likely relies heavily on syntactic features for genre prediction especially in these languages, as lexical information is less informative due to the lack of vocabulary overlap between the training and test sets. In contrast, for texts in Latin script, the model might be able to mitigate the effects of syntactical perturbations caused by the adversarial attack by leveraging shared vocabulary. As illustrated in Figure 9, the performance decline for test sets in Latin scripts, resulting from the shuffled word order, is comparable to or slightly greater than the decline observed when half of the words are removed. The translation to English with the machine translation system was demonstrated to be detrimental to the model's performance as well. The "translate-and-test" pipeline led to a decrease in macro-F1 scores ranging from seven points (Catalan test set) to more than sixteen points (Greek, Ukrainian, and Slovenian).

9. Conclusion

In the study, we assess the zero-shot cross-lingual capabilities of the Transformer-based X-GENRE classifier (Kuzman et al. 2023a) on diverse European languages. Through experiments conducted on a novel manually-annotated test dataset comprising ten languages from three language families, we demonstrate that the X-GENRE classifier is capable of zero-shot cross-lingual transfer across diverse languages. Notably, the classifier achieves macro-F1 scores ranging from 0.80 to 0.95 on all languages except Maltese. These findings indicate that the openly-available X-GENRE classifier can be applied on corpora in any of the 100 languages it was pretrained on for high-quality automatic enrichment with genre information.

To ascertain the potential efficiency of the genre classifier on additional languages not included in the test dataset, we analyse the impact of various factors on its performance. These factors include the degree of linguistic relatedness between the test language and the training language, the use of non-Latin scripts, and the extent of the coverage of the language in the XLM-RoBERTa (Conneau et al. 2020) pretraining dataset, since the X-GENRE classifier is a fine-tuned version of this massively multilingual model. The experiments lead to the following findings:

- **Degree of Linguistic Relatedness** The study reveals that the degree of linguistic relatedness between the test language and the training language does not significantly impact cross-lingual performance. Inter alia, the model exhibits superior performance on Turkish, a language from a different language family than the training languages (English and Slovenian), compared to Croatian, which is closely related to Slovenian in the training data.
- **Performance Across Scripts** The model performance is not negatively impacted when applied across scripts. Even though the X-GENRE model was fine-tuned on texts in Latin script, it demonstrates strong performance on languages that use Cyrillic or Greek script. As test sets in non-Latin scripts have a limited vocabulary overlap with the Latin training dataset, we extend the analyses to assess the impact of vocabulary overlap on the model's performance. An analysis of the token overlap between the training and test datasets reveals that the model's performance is not dependent on the presence of shared vocabulary.
- **Test Language Coverage in the Pretraining Dataset** The analysis of the relationship between the coverage of the test language within the XLM-RoBERTa pretraining data and the performance of the model on the test set indicates that even low coverage (three to five gigabytes of texts) of the test language in the pretraining data suffices for the X-GENRE model to exhibit strong performance. However, in cases where the test language is entirely absent from the pretraining data, such as Maltese, achieving satisfactory cross-lingual performance is unlikely. Interestingly, substantial coverage of a test language in the pretraining data might have a positive impact on cross-lingual performance, as evidenced by the classifier achieving the highest macro-F1 scores on Ukrainian, the language with the highest coverage in the pretraining dataset.
- **Impact of Syntax versus Vocabulary** The examination of the effects of two adversarial attacks, specifically removing half of the words from the text and shuffling the word order, on the cross-lingual performance of the model demonstrates that shuffling the word order, which disrupts syntax, has greater or comparable influence on performance compared to the removal of vocabulary. This finding highlights the significant role of syntax in cross-lingual genre identification.
- **Impact of Machine Translation on Performance** Translating the texts into English with a machine translation system and applying the classifier on the translated text is shown to provide inferior results than the zero-shot cross-lingual classification. Translation is found to particularly hinder the identification of genres that are characterized by distinct language styles, such as *Forum* that is known for its non-standard language. These findings highlight

the suitability of zero-shot cross-lingual classification as the preferred approach for genre classification in languages lacking language-specific genre classifiers.

The newly developed manually-annotated X-GINCO test dataset in ten languages is available upon request from the corresponding author. Additionally, to encourage further research, we have integrated the dataset into the AGILE Benchmark (Automatic Genre Identification Benchmark), which can be accessed at GitHub.[k] The test dataset is intentionally not published online to prevent its inclusion in the training data of large language models (LLMs), which ensures reliable evaluation of future LLMs and other technologies.

Motivated by positive results, as an additional contribution, we publish the MaCoCu corpora automatically annotated with genres under the name "MaCoCu-Genre" datasets. The newly available dataset collection encompasses approximately 52 million texts in nine test languages (excluding Maltese), enriched with genre information.[l] We have not published the Maltese web corpus, as cross-lingual genre prediction for this language did not yield satisfactory results.

The MaCoCu-Genre corpus collection has now already been used in several studies which leveraged the provided genre information to facilitate a more controlled development of training and test datasets. Specifically, in a study investigating the development of training data for news topic classification using large language models, a collection of 20,000 news texts was extracted from the Slovenian, Croatian, Catalan, and Greek MaCoCu-Genre corpora by filtering the datasets based on the *News* genre label (Kuzman and Ljubešić 2025). Another study used the Slovenian genre-enriched corpus to develop a question-answering test dataset for the evaluation of retrieval-augmented generation systems in the Slovenian language (Kuzman et al. 2024). The authors noted that, due to the high level of comparability of the MaCoCu-Genre corpora in terms of their construction methodology and format, the same approach can be readily applied to other corpora within the collection to develop question-answering datasets for additional languages.

In further work, we plan to extend our cross-lingual analyses to additional languages. The selection of languages for the current study is based on the availability of the languages in the MaCoCu corpora (Bañón et al. 2022) to ensure high comparability of the test sets. Consequently, the study is limited to a selection of European languages. Moving forward, we aim to extend our evaluation to languages from diverse regions around the world.

Furthermore, the analysis reveals that the model achieves inferior performance on Maltese due to the absence of this language in the XLM-RoBERTa pretraining data. Previous work (Muller et al. 2021; Chau et al. 2020; Ebrahimi and Kann 2021) has demonstrated that additional pretraining of the XLM-RoBERTa model with monolingual data notably improves the performance of massively multilingual BERT-like models on Maltese across various tasks. Future research could explore using a model that is additionally pretrained on Maltese and other languages excluded from the XLM-RoBERTa pretraining dataset and fine-tuning it with the English-Slovenian X-GENRE dataset. Another promising direction would be to investigate whether fine-tuning the XLM-RoBERTa model on a manually annotated genre training dataset in a language that is not included in the XLM-RoBERTa pretraining dataset would improve its performance on this language, or if additional pretraining is critical. An intriguing question to address in this context is the balance between pretraining and fine-tuning: to what extent does additional pretraining solve language-specific challenges compared to direct fine-tuning on a target language dataset without prior pretraining? These approaches could yield valuable insights into optimizing multilingual models for low-resource languages.

[k] `https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark`

[l] The genre-enriched MaCoCu-Genre web corpora are available in the CLARIN.SI repository at `http://hdl.handle.net/11356/1969`.

Competing Interests

The authors declare no competing interests.

Acknowledgements. The research presented in this paper was conducted within the research project "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642), the CLARIN.SI research infrastructure, and the research programme "Language resources and technologies for Slovene" (P6-0411), all funded by the Slovenian Research and Innovation Agency (ARIS).

References

Agrawal, S., **Sanagavarapu, L. M.**, **and Reddy, Y. R.** (2019). FACT-Fine grained Assessment of web page CredibiliTy. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pp. 1088–1097. IEEE.

Alves, D., **Bekavac, B.**, **Zeman, D.**, **and Tadić, M.** (2023). Analysis of corpus-based word-order typological methods. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pp. 36–46.

Amjad, M., **Sidorov, G.**, **and Zhila, A.** (2020). Data augmentation using machine translation for fake news detection in the Urdu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 2537–2542.

Bañón, M., **Esplà-Gomis, M.**, **Forcada, M. L.**, **García-Romero, C.**, **Kuzman, T.**, **Ljubešić, N.**, **van Noord, R.**, **Sempere, L. P.**, **Ramírez-Sánchez, G.**, **Rupnik, P.**, **and others** (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation*, pp. 301–302.

Baroni, M., **Bernardini, S.**, **Ferraresi, A.**, **and Zanchetta, E.** (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Boese, E. S. (2005). *Stereotyping the web: genre classification of web documents*. PhD thesis, Citeseer.

Chau, E. C., **Lin, L. H.**, **and Smith, N. A.** (2020). Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1324–1334.

Conneau, A., **Khandelwal, K.**, **Goyal, N.**, **Chaudhary, V.**, **Wenzek, G.**, **Guzmán, F.**, **Grave, É.**, **Ott, M.**, **Zettlemoyer, L.**, **and Stoyanov, V.** (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451.

de Vries, W., **Wieling, M.**, **and Nissim, M.** (2022). Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *The 60th Annual Meeting of the Association for Computational Linguistics*, pp. 7676–7685. Association for Computational Linguistics (ACL).

Dryer, M. S. and Haspelmath, M. (2013). The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. *Online: http://wals. info*.

Eberhard, D. M., **Simons, G. F.**, **and Fenning, C. D.** (2023). Ethnologue: Languages of the World. Twenty-sixth edition.

Ebrahimi, A. and Kann, K. (2021). How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4555–4567.

Egbert, J., **Biber, D.**, **and Davies, M.** (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Eskelinen, A., **Myntti, A.**, **Henriksson, E.**, **Pyysalo, S.**, **and Laippala, V.** (2024). Building question-answer data using web register identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2595–2611.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In *fifth Web as Corpus workshop*, pp. 27–35.

Henriksson, E., **Myntti, A.**, **Eskelinen, A.**, **Erten-Johansson, S.**, **Hellström, S.**, **and Laippala, V.** (2024). Untangling the unrestricted web: Automatic identification of multilingual registers. *arXiv preprint arXiv:2406.19892*.

Hu, J., **Ruder, S.**, **Siddhant, A.**, **Neubig, G.**, **Firat, O.**, **and Johnson, M.** (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pp. 4411–4421. PMLR.

Joulin, A., **Grave, É.**, **Bojanowski, P.**, **and Mikolov, T.** (2017). Bag of Tricks for Efficient Text Classification. In *15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431.

Karthikeyan, K., **Wang, Z.**, **Mayhew, S.**, **and Roth, D.** (2019). Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*.

Khalil, T., **Kiełczewski, K.**, **Chouliaras, G. C.**, **Keldibek, A.**, **and Versteegh, M.** (2019). Cross-lingual intent classification

in a low resource industrial setting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6419–6424.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kuzman, T. and Ljubešić, N. (2022). Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In **Mladenić, D. and Grobelnik, M.**, editors, *Data Mining and Data Warehouses - SiKDD, 10 October 2022, Ljubljana, Slovenia*. Jožef Stefan Institute.

Kuzman, T. and Ljubešić, N. (2023). Automatic genre identification: a survey. *Language Resources and Evaluation*, pp. 1–34.

Kuzman, T. and Ljubešić, N. (2024)a. English-Slovenian text genre dataset X-GENRE. Slovenian language resource repository CLARIN.SI.

Kuzman, T. and Ljubešić, N. (2024)b. Multilingual text genre classification model x-GENRE. Slovenian language resource repository CLARIN.SI.

Kuzman, T., **Ljubešić, N.**, **and Pollak, S.** (2022)a. Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In **Fišer, D. and Erjavec, T.**, editors, *Proceedings of the Conference on Language Technologies and Digital Humanities*, 100–107. Institute of Contemporary History.

Kuzman, T. and Ljubešić, N. (2025). LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, 13:35621–35633.

Kuzman, T., **Mozetič, I.**, **and Ljubešić, N.** (2023)a. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Kuzman, T., **Pavleska, T.**, **Rupnik, U.**, **and Cigoj, P.** (2024). PandaChat-RAG: Towards the benchmark for Slovenian RAG Applications. In *Proceedings of the Slovenian Conference on Artificial Intelligence 2024, Vol. A, Ljubljana, Slovenia*, 7–10. Institute „Jožef Stefan“.

Kuzman, T., **Rupnik, P.**, **and Ljubešić, N.** (2022)b. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In *Language Resources and Evaluation Conference*, pp. 1584–1594, Marseille, France. European Language Resources Association.

Kuzman, T., **Rupnik, P.**, **and Ljubešić, N.** (2023)b. Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pp. 91–103.

Laippala, V., **Egbert, J.**, **Biber, D.**, **and Kyröläinen, A.-J.** (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*, pp. 1–32.

Laippala, V., **Kyllönen, R.**, **Egbert, J.**, **Biber, D.**, **and Pyysalo, S.** (2019). Toward multilingual identification of online registers. In *22nd Nordic Conference on Computational Linguistics*, pp. 292–297.

Laippala, V., **Rönnqvist, S.**, **Hellström, S.**, **Luotolahti, J.**, **Repo, L.**, **Salmela, A.**, **Skantsi, V.**, **and Pyysalo, S.** (2020). From web crawl to clean register-annotated corpora. In *12th Web as Corpus Workshop*, pp. 14–22.

Laippala, V., **Salmela, A.**, **Rönnqvist, S.**, **Aji, A. F.**, **Chang, L.-H.**, **Dhifallah, A.**, **Goulart, L.**, **Kortelainen, H.**, **Pàmies, M.**, **Dutra, D. P.**, **and others** (2022). Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pp. 215–221.

Lauscher, A., **Ravishankar, V.**, **Vulić, I.**, **and Glavaš, G.** (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499.

Lepekhin, M. and Sharoff, S. (2022). Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for the Genre Classification Task. In *Language Resources and Evaluation Conference*, pp. 5974–5982, Marseille, France. European Language Resources Association.

Lin, Y.-H., **Chen, C.-Y.**, **Lee, J.**, **Li, Z.**, **Zhang, Y.**, **Xia, M.**, **Rijhwani, S.**, **He, J.**, **Zhang, Z.**, **Ma, X.**, **and others** (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3125–3135.

Littell, P., **Mortensen, D. R.**, **Lin, K.**, **Kairis, K.**, **Turner, C.**, **and Levin, L.** (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14.

Ljubešić, N. and Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3271–3282.

Martínez-García, A., **Badia, T.**, **and Barnes, J.** (2021). Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3136–3153.

Muller, B., **Anastasopoulos, A.**, **Sagot, B.**, **and Seddah, D.** (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 448–462.

Müller-Eberstein, M., **van der Goot, R.**, **and Plank, B.** (2021). Genre as Weak Supervision for Cross-lingual Dependency Parsing. In *2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4786–4802.

OpenAI (2023). GPT-4 Technical Report.

Orlikowski, W. J. and Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pp. 541–574.

Ortiz Suárez, P. J., **Sagot, B.**, **and Romary, L.** (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Philippy, F., **Guo, S.**, **and Haddadan, S.** (2023). Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*.

Pires, T., **Schlinger, E.**, **and Garrette, D.** (2019). How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.

Priyatam, P. N., **Iyengar, S.**, **Perumal, K.**, **and Varma, V.** (2013). Don't Use a Lot When Little Will Do: Genre Identification Using URLs. *Res. Comput. Sci.*, 70:233–243.

Repo, L., **Skantsi, V.**, **Rönnqvist, S.**, **Hellström, S.**, **Oinonen, M.**, **Salmela, A.**, **Biber, D.**, **Egbert, J.**, **Pyysalo, S.**, **and Laippala, V.** (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, EACL 2021*, pp. 183–191. Association for Computational Linguistics (ACL).

Rönnqvist, S., **Skantsi, V.**, **Oinonen, M.**, **and Laippala, V.** (2021). Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification. In *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 157–165.

Roussinov, D., **Crowston, K.**, **Nilan, M.**, **Kwasnik, B.**, **Cai, J.**, **and Liu, X.** (2001). Genre based navigation on the web. In *34th annual Hawaii international conference on system sciences*, pp. 10–pp. IEEE.

Sharoff, S. (2010). In the Garden and in the Jungle. In *Genres on the Web*, pp. 149–166. Springer.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*.

Skantsi, V. and Laippala, V. (2023). Analyzing the unrestricted web: The Finnish corpus of online registers. *Nordic Journal of Linguistics*, pp. 1–31.

Stein, B., **Eissen, S. M. z.**, **and Lipka, N.** (2010). Web genre analysis: Use cases, retrieval models, and implementation issues. In *Genres on the Web*, pp. 167–189. Springer.

Stewart, J. G. and Callan, J. (2009). *Genre oriented summarization*. PhD thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. *Proc. Towards a Reference Corpus of Web Genres*.

Tan, S. and Monz, C. (2023). Towards a better understanding of variations in zero-shot neural machine translation performance. *arXiv preprint arXiv:2310.10385*.

Tebbifakhr, A., **Bentivogli, L.**, **Negri, M.**, **and Turchi, M.** (2019). Machine Translation for Machines: the Sentiment Classification Use Case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1368–1374.

Ulčar, M., **Žagar, A.**, **Armendariz, C. S.**, **Repar, A.**, **Pollak, S.**, **Purver, M.**, **and Robnik-Šikonja, M.** (2021). Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

Unanue, I. J., **Haffari, G.**, **and Piccardi, M.** (2023). T3l: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics*, 11:1147–1161.

Van der Wees, M., **Bisazza, A.**, **and Monz, C.** (2018). Evaluation of machine translation performance across multiple genres and languages. In *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

van Noord, R., **Kuzman, T.**, **Rupnik, P.**, **Ljubešić, N.**, **Esplà-Gomis, M.**, **Ramírez-Sánchez, G.**, **and Toral, A.** (2024). Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across 11 Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5221–5234.

Vidulin, V., **Luštrek, M.**, **and Gams, M.** (2007). Using genres to improve search engines. In *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pp. 45–51.

Wenzek, G., **Lachaux, M.-A.**, **Conneau, A.**, **Chaudhary, V.**, **Guzmán, F.**, **Joulin, A.**, **and Grave, E.** (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France. European Language Resources Association.

Xue, L., **Constant, N.**, **Roberts, A.**, **Kale, M.**, **Al-Rfou, R.**, **Siddhant, A.**, **Barua, A.**, **and Raffel, C.** (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online. Association for Computational Linguistics.

Zu Eissen, S. M. and Stein, B. (2004). Genre classification of web pages. In *Annual Conference on Artificial Intelligence*, pp. 256–269. Springer.

4.7 Related Paper: State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?

State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?

Taja Kuzman Pungeršek*, Peter Rupnik*, Ivan Porupski*, Vuk Dinić*, Nikola Ljubešić*†‡
*Jožef Stefan Institute;
†Faculty of Computer and Information Science, University of Ljubljana;
‡Institute of Contemporary History;
Ljubljana, Slovenia
{taja.kuzman, peter.rupnik, ivan.porupski, vuk.dinic, nikola.ljubesic}@ijs.si

Abstract

Until recently, fine-tuned BERT-like models provided state-of-the-art performance on text classification tasks. With the rise of instruction-tuned decoder-only models, commonly known as large language models (LLMs), the field has increasingly moved toward zero-shot and few-shot prompting. However, the performance of LLMs on text classification, particularly on less-resourced languages, remains under-explored. In this paper, we evaluate the performance of current language models on text classification tasks across several South Slavic languages. We compare openly available fine-tuned BERT-like models with a selection of open-weight and closed-source LLMs across three tasks in three domains: sentiment classification in parliamentary speeches, topic classification in news articles and parliamentary speeches, and genre identification in web texts. Our results show that LLMs demonstrate strong zero-shot performance, often matching or surpassing fine-tuned BERT-like models. Moreover, when used in a zero-shot setup, LLMs perform comparably in South Slavic languages and English. However, we also point out key drawbacks of LLMs, including less predictable outputs, significantly slower inference, and higher computational costs. Due to these limitations, fine-tuned BERT-like models remain a more practical choice for large-scale automatic text annotation.

Keywords: LLM evaluation, text classification, large language models, South Slavic languages, sentiment identification, topic classification, genre identification

1. Introduction

Until recently, the dominant approach for text classification tasks relied on fine-tuning BERT-like transformer models on thousands of manually-annotated training examples. Recently, however, the field has shifted with the development of instruction-tuned decoder-only transformer models. These models, also commonly referred to as large language models (LLMs), which were originally developed primarily for text generation tasks, have demonstrated remarkable capabilities across a broad range of natural language processing (NLP) tasks, including text classification (Kuzman et al., 2023; Huang et al., 2023).

In this paper, we focus on South Slavic languages, where research on text classification tasks included in our study has, until recently, been limited or even non-existent (Kuzman and Ljubešić, 2023; Mochtak et al., 2024; Kuzman and Ljubešić, 2025). We take a first step toward systematically evaluating the current state of the art for text classification in these languages. Our evaluation is based on three text classification tasks in three different domains for which manually-annotated test datasets in South Slavic languages and fine-tuned BERT-like classifiers are freely available: sentiment classification of parliamentary speeches, topic classification in news articles, topic classification in parliamentary speeches, and automatic genre identification in web texts. These tasks span different domains and language styles, allowing for a comprehensive analysis of the performance of transformer-based models on text classification tasks. Specifically, we compare the performance of openly available fine-tuned BERT-like models with the zero-shot capabilities of both open-weight and closed-source LLMs used via prompting.

An important aspect of our study is to examine whether the performance of multilingual models on South Slavic languages is on par with their performance on English. This question is particularly relevant given that the evaluated large language models have been predominantly pretrained and instruction-tuned on English data.

By evaluating various models on a selection of text classification tasks in English and various South Slavic languages, we set out to test the following two hypotheses that are based on previous experiments with fine-tuned BERT-like models and LLMs on automatic genre identification (Kuzman et al., 2023), news topic classification (Kuzman and Ljubešić, 2025) and sentiment analysis in parliamentary texts (Mochtak et al., 2025):

H1 Zero-shot prompting with instruction-tuned large language models (LLMs) can achieve

Dataset	Lang	# Instances	# Labels	% Most and Least Frequent Label
Sentiment classification in parliamentary speeches				
ParlaSent-EN-test	EN	2600	3	40.8% (Neutral), 26.8% (Positive)
ParlaSent-HR-test	HR	1336	3	41.9% (Negative), 17.2% (Positive)
ParlaSent-SR-test	SR	1074	3	46.2% (Negative), 17.6% (Positive)
ParlaSent-BS-test	BS	190	3	47.9% (Negative), 14.7% (Positive)
Genre classification in web texts				
EN-GINCO	EN	272	8	23.5% (Information/Explanation), 0.4% (Legal)
X-GINCO-SL	SL	80	8	15% (Prose/Lyrical), 8.8% (Opinion/Argumentation)
X-GINCO-HR	HR	80	8	16.3% (Promotion), 7.5% (Instruction)
X-GINCO-MK	MK	80	8	15% (News), 1% (Opinion/Argumentation)
Topic classification in news articles				
IPTC-test-HR	HR	291	17	11.0% (Economy), 3.8% (Conflict, War and Peace)
IPTC-test-SL	SL	282	17	10.6% (Society), 3.2% (Conflict, War and Peace)
Topic classification in parliamentary speeches				
ParlaCAP-test-EN	EN	876	22	6.4% (Law and Crime), 2.1% (Culture)
ParlaCAP-test-HR	HR	869	22	8.5% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-SR	SR	874	22	7.1% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-BS	BS	824	22	10.4% (Other), 0.5% (Culture)

Table 1: Information on test datasets in English (EN), Croatian (HR), Serbian (SR), Bosnian (BS), Slovenian (SL), and Macedonian (MK).

results comparable to the use of BERT-like models fine-tuned on training data that are similar to the test data.

H2 The performance of LLMs used in a zero-shot setup on text classification tasks on South Slavic test datasets is comparable to the performance on English test datasets.

2. Related Work

After the introduction of transformer architectures, BERT (bidirectional encoder representations from transformers) models have achieved state-of-the-art results in text classification tasks, outperforming earlier non-neural approaches, such as support vector machines (SVMs). They have also demonstrated strong cross-lingual zero-shot capabilities in various classification tasks, including automatic genre identification (Kuzman and Ljubešić, 2023), news topic classification (Petukhova and Fachada, 2023; De Clercq et al., 2020), and sentiment classification (Mochtak et al., 2024). However, these models still require fine-tuning on a training dataset, developed during manual annotation campaigns that are time-consuming and costly.

Instruction-tuned decoder-only transformer models, commonly referred to as large language models (LLMs), have recently shown strong performance in a range of classification tasks, even in zero-shot prompting setups that require no training data (Kuzman et al., 2023; Ljubešić et al., 2024a; Huang et al., 2023; Kuzman Pungeršek et al., 2026). They have achieved promising results on various natural language processing tasks, including stance detection (Zhang et al., 2022), implicit hate speech categorization (Huang et al., 2023), news topic classification (Kuzman and Ljubešić, 2025), automatic genre identification (Kuzman et al., 2023), causal commonsense reasoning (Ljubešić et al., 2024b), and machine translation (Hendy et al., 2023). Due to their promising performance, researchers have even started using them as data annotators, either by generating text and labels (Meng et al., 2022) or by annotating pre-existing texts (Kuzman and Ljubešić, 2025; Kuzman Pungeršek et al., 2026). Despite the growing interest in this topic, the majority of evaluations of LLMs used in

text classification tasks are limited only to English (Sun et al., 2023; Zhang et al., 2025; Kostina et al., 2025; Zhao et al., 2024). Systematic multilingual evaluations, especially which would include less-resourced languages such as those in the South Slavic group remain limited. Our work addresses this gap by providing a comparative evaluation of open-weight and closed-source LLMs with openly-available fine-tuned BERT-like models across four benchmark families comprising three diverse classification tasks and three different domains in South Slavic languages and English.

3. Benchmarks

The benchmarks (evaluation datasets) used in this study cover three text classification tasks, namely, sentiment identification, topic classification, and automatic genre identification, and three domains: parliamentary speeches, news articles and web texts. An overview of the datasets is provided in Table 1. The four benchmark families differ significantly in terms of language coverage, number of test instances, and label granularity.

The topic classification task is evaluated on two domains: 1) news articles, namely, the Croatian and Slovenian IPTC test datasets (Kuzman and Ljubešić, 2025), which comprise around 300 text instances per language, and 2) parliamentary speeches, namely, the Bosnian, Croatian, English and Serbian ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) that consist of approximately 820 to 880 instances per language. In the ParlaCAP benchmarks, an instance is a transcription of an utterance given by a parliamentary member in a parliamentary session.

The topic classification task involves the highest number of labels, that is, 17 news topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema[1] (IPTC, 2022), and 22 policy topic labels (21 major topics and a label *Other*) from the Comparative Agendas Project (CAP; Baumgartner et al., 2019) Master Codebook (Bevan, 2019).[2]

In contrast, the Bosnian, Croatian, English, and Serbian ParlaSent sentiment identification datasets (Mochtak et al., 2024; Mochtak et al., 2023) have a significantly lower granularity of labels, with only 3 categories. They are represented by the largest number of instances, ranging from 190 (Bosnian part) to 2600 (English part) sentence-level instances.

With 8 labels, the Croatian, English, Macedonian, and Slovenian GINCO genre datasets (Kuzman et al., 2023) represent a midpoint in label granularity among the four benchmark families. However, the genre identification task might be the most difficult one, as genre identification depends on the interpretation of full texts with the focus on author's purpose, the common function of the text, and the text's conventional form (Orlikowski and Yates, 1994). This complexity has also contributed to smaller test datasets in terms of the number of text instances, as manual annotation is more time-consuming. It is also important to note that, unlike the parliamentary datasets, the English portion of the genre datasets is not fully comparable to the South Slavic portions, which are label-balanced and contain fewer ambiguous instances. Nevertheless, the genre datasets remain valuable for evaluating model performance within each language.

All test datasets were manually annotated by annotators that are deemed reliable based on their satisfactory inter-annotator agreement, namely, Krippendorff's alpha (Krippendorff, 2018) values close to or above the 0.667 threshold for reliable annotation. To prevent large language models from incorporating the test datasets during their training phase, the test datasets are not publicly available, except for the ParlaSent benchmark family. Access to other datasets is granted on request from the corresponding authors. Further details on the test datasets are provided in Section A.1 of the Appendix.

4. Methodology

In this paper, we evaluate the main machine learning approaches that have recently been used for our selection of text classification tasks, with a focus on the comparison between the freely available fine-tuned BERT models and the open-weight and closed-source LLMs.[3] The models are evaluated on four families of test datasets that comprise South Slavic languages. The performance of the models is evaluated based on the micro-F1 and macro-F1 metrics, which enable assessment of the model performance at both the instance and label levels, respectively.

The following machine learning models are included in the evaluation:

- **dummy classifier**: a dummy classifier that predicts the most frequent class in the training data. To allow comparison, the dummy classifiers were trained on the same datasets that were used for fine-tuning the BERT-like models, mentioned below.

[1] https://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

[2] https://www.comparativeagendas.net/pages/master-codebook

[3] The code for the model evaluation and analysis of results is available at https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic.

- **fine-tuned BERT-like classifiers**: in our study, we evaluate previously developed openly accessible multilingual fine-tuned BERT-like models that have been fine-tuned for the respective task, namely, the XLM-R-ParlaSent (Rupnik et al., 2023; Mochtak et al., 2024) model for sentiment identification in parliamentary texts, the X-GENRE classifier (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d, 2023) for automatic genre identification, the IPTC News Topic classifier (Kuzman and Ljubešić, 2025; Kuzman and Ljubešić, 2025) for news topic classification, and the ParlaCAP classifier (Kuzman Pungeršek et al., 2026; Kuzman Pungeršek and Ljubešić, 2025) for topic classification in parliamentary speeches. The XLM-R-ParlaSent and the ParlaCAP models are based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on parliamentary proceedings in 30 European languages (Mochtak et al., 2024). The XLM-R-ParlaSent model was fine-tuned on 13 thousand instances from the ParlaSent sentiment training dataset (Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian; Mochtak et al., 2024), while the ParlaCAP model was fine-tuned on the ParlaCAP-train dataset (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026) that comprises around 30 thousand speeches from parliamentary debates annotated with CAP topic labels, originating from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The X-GENRE classifier is based on the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a) in English and Slovenian; while the IPTC News Topic classifier is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) that was fine-tuned on the EMMediaTopic dataset (Kuzman and Ljubešić, 2024c) in Catalan, Croatian, Greek, and Slovenian. All fine-tuned models use the same classes as the test datasets used in our study.

- **open-weight and closed-source large language models**: we use closed-source OpenAI models, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`; OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`; OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`; OpenAI, 2025); a closed-source Gemini 2.5 Flash model (Comanici et al., 2025) by Google DeepMind; a closed-source Mistral Medium 3.1 model (`mistral-medium-2508`; Mistral AI, 2025) by Mistral AI; and four open-weight models, namely, the Meta LLaMA 3.3 model (Meta, 2024), the Gemma 3 model (Gemma Team et al., 2025), the Qwen 3 model (Yang et al., 2025), and the DeepSeek-R1-Distill model (`DeepSeek-R1-Distill-Qwen-14B`; Guo et al., 2025). It is important to note that while the LLaMA model was pretrained on a web text collection in various languages, it is said to support only 8 languages, namely English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai (Meta, 2024). The DeepSeek-R1-Distill model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that provides support for more than 29 languages – not including South Slavic languages though. In contrast, the Gemma 3 model is reported to support over 140 languages (Gemma Team et al., 2025), and the Qwen 3 model was pretrained on 119 languages (Yang et al., 2025). While closed-source models are said to be massively multilingual, with Gemini 2.5 models being pretrained on over 400 languages (Comanici et al., 2025), details on their language coverage are very limited.

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). OpenAI models were used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[4] that provides a unified API access to various closed-source models. To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. More details on the models and their implementation, including information on the availability of openly available models and fine-tuning datasets, are provided in Section A.2 of the Appendix.

All instruction-tuned LLMs are used in a zero-shot prompting setup, meaning that they receive only a task description and label definitions. All prompts and label definitions are written in English, while the instances are provided in the original languages (English or South Slavic). Changing the language of the prompt could introduce an additional factor that affects performance. In the presented experiments, our goal is to assess model performance on a specific task and language, rather than to evaluate their instruction-following abilities across different languages. The models are instructed to output a label, represented by a digit. The same prompt per benchmark family is used for all LLMs.

[4] `https://openrouter.ai/`

Prompts are provided in Figure 4 in Section A.2 of the Appendix.

5. Results

In this section, we evaluate the performance of the fine-tuned BERT-like models and the instruction-tuned LLMs on a selection of text classification tasks that include test datasets in South Slavic languages. First, in Section 5.1, we provide results on the four benchmark families with a focus on hypothesis H1, which expects that zero-shot prompting with LLMs can provide performance that is comparable to that of fine-tuned BERT-like models. In Section 5.2, we compare in more detail the performance of the closed-source and open-weight LLMs on the three text classification tasks, which is followed by a discussion on the advantages and limitations of LLMs for data annotation based on text classification tasks (Section 5.3). Lastly, in Section 5.4, we compare the performance of LLMs on English test datasets with their performance on South Slavic datasets, addressing hypothesis H2, which presumes that the available multilingual LLMs perform similarly on South Slavic languages as on English.

Model	Rank	Rank (EN)	Rank (South Slavic)
GPT-5	2.29	1.33	2.55
GPT-4o	2.36	2.00	2.45
Fine-Tuned BERT-Like Model	3.21	4.67	2.82
Gemini 2.5 Flash	3.50	3.33	3.55
Mistral Medium 3.1	5.36	5.00	5.45
Gemma 3	5.71	5.67	5.73
LLaMA 3.3	6.00	6.67	5.82
Qwen 3	7.43	7.00	7.55
GPT-3.5-Turbo	8.79	9.00	8.73
DeepSeek-R1-Distill	10.00	10.00	10.00

Table 2: Comparison of models based on their average rank (1 = best-performing, 10 = worst-performing) across all test datasets (first column), and averaged across English (second column) or South Slavic (third column) test datasets.

5.1. State of the Art in Text Classification Tasks

Figure 1 provides the results of model evaluation on our selection of text classification tasks. A consistent pattern emerges across all four benchmark families: LLMs, when used in a zero-shot prompting setup, achieve some of the highest scores. As shown in Table 2, which compares model rankings across tasks, LLMs achieve first place more often on average than the fine-tuned BERT-like model.

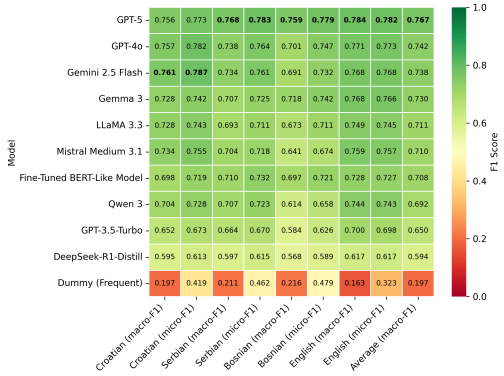
Figure 1a shows that both open-weight and closed-source LLMs, used in a zero-shot prompting setup on the sentiment identification task, achieve performance that is comparable or even significantly higher than that of a fine-tuned BERT-like model trained on a large manually-annotated sentiment dataset. The only models that consistently perform worse than the fine-tuned BERT-like model are GPT-3.5-Turbo and DeepSeek-R1-Distill. Sentiment classification appears broad enough that more potent LLMs can interpret label definitions effectively without task-specific fine-tuning, reducing the benefit of additional training.

In contrast, fine-tuned BERT-like models outperform most LLMs on automatic genre identification and topic classification tasks. These tasks depend on predefined label sets based on specific guidelines, and the strong performance of fine-tuned BERT-like models indicates that domain-specific fine-tuning on labelled data still offers an advantage over the general knowledge leveraged by LLMs in zero-shot setups. This advantage is particularly clear in genre identification for South Slavic texts, where the fine-tuned BERT-like model significantly outperforms LLMs. The likely reason for the fine-tuned model's very strong performance on South Slavic genre datasets is the curated nature of the test data – more challenging examples were removed before and during manual annotation, unlike in the English genre test dataset where the instances were randomly sampled from an English web corpus. Nevertheless, despite this limitation, the South Slavic test dataset remains valuable for comparing the performance of LLMs.

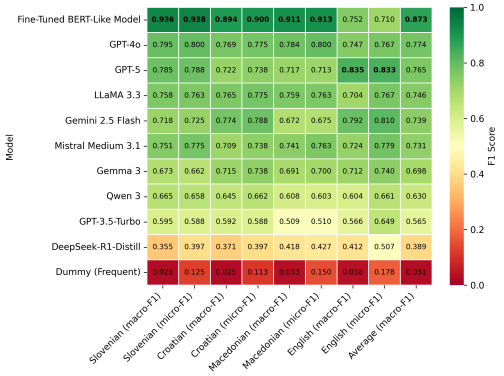
To conclude, since some LLMs used in a zero-shot prompting setup achieve higher or comparable results to fine-tuned BERT-like models across all classification tasks and languages, as shown in Table 2, we can confirm hypothesis H1, which proposed that zero-shot prompting with LLMs can perform comparably to fine-tuned BERT-like models.

5.2. Comparison of Large Language Models

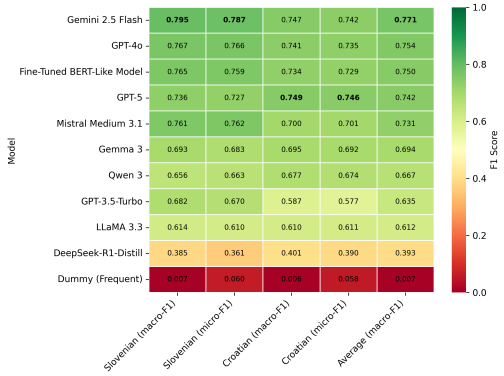
Figure 2 shows the performance of open-weight and closed-source LLMs, used via prompting, on the tasks of sentiment identification, automatic genre identification, news topic classification, and parliamentary topic classification. The DeepSeek-R1-Distill model is not included in the comparison, as it performs significantly worse than the other



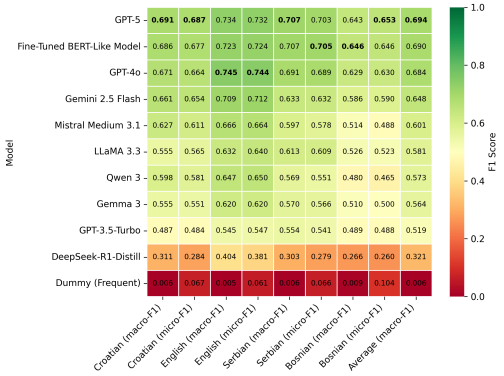
(a) Sentiment classification.



(b) Automatic genre identification.



(c) News topic classification.



(d) Parliamentary topic classification.

Figure 1: Micro-F1 and macro-F1 scores across models and languages on the test datasets for sentiment classification (Figure 1a), automatic genre identification (Figure 1b), and topic classification on news (Figure 1c) and parliamentary speeches (Figure 1d).

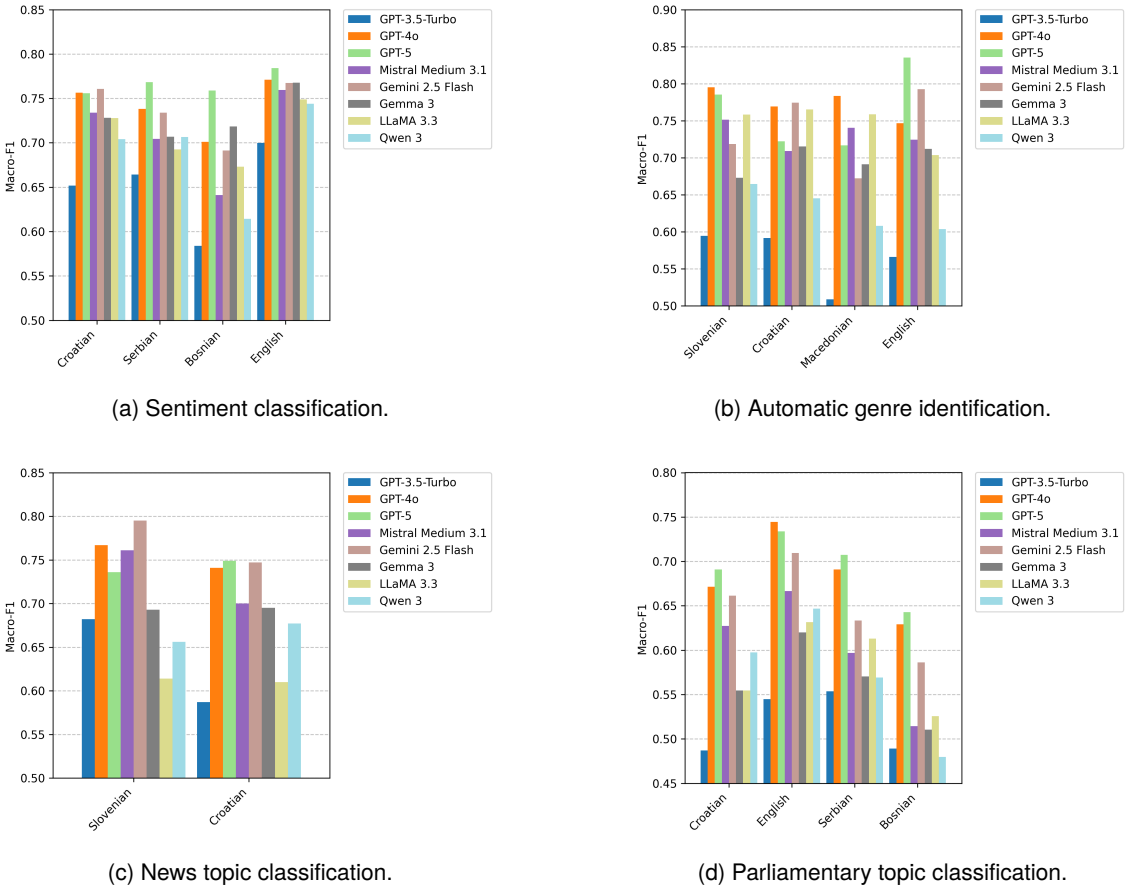
models, as shown in Figure 1.

While different models perform best across different languages and test datasets, a clear trend emerges: the top-performing models across all four benchmark families are the closed-source GPT-4o and GPT-5 from OpenAI, along with Gemini 2.5 Flash. Although GPT-5 is newer and reportedly more powerful, it does not outperform GPT-4o on all benchmarks. Among open-weight models, Gemma 3 generally achieves the best results in sentiment identification (Figure 2a) and news topic classification (Figure 2c). For automatic genre identification (Figure 2b) and parliamentary topic classification (Figure 2d), rankings of open-weight models vary by language. Overall, the weakest performance is observed with the older closed-source GPT-3.5-Turbo model, highlighting the rapid progress in both open-weight and closed-source model development.

5.3. Advantages and Disadvantages of LLMs

A clear advantage of LLMs is that they do not require manually-annotated training data for specific tasks, yet still achieve strong performance when provided only with task instructions and brief label descriptions. However, these models are significantly more computationally expensive than fine-tuned BERT-like models. While closed-source models deliver the best performance, as shown in previous sections, they come with several limitations: they are costly to use, their architectures and pre-training data are not publicly disclosed, and access through APIs hinders reproducibility, in contrast to open-weight LLMs and fine-tuned BERT-like models.

What is more, the inference speed of all LLMs is significantly slower than that of a fine-tuned BERT-like model. As shown in Figure 3, the fine-tuned BERT-like model achieves one of the highest macro-F1 scores on the topic classification task for parliamentary speeches, while maintaining a very low inference time of just 0.02 seconds per



(a) Sentiment classification.

(b) Automatic genre identification.

(c) News topic classification.

(d) Parliamentary topic classification.

Figure 2: Comparison of LLMs used in a zero-shot prompting setup on sentiment identification (Figure 2a), automatic genre identification (Figure 2b), and topic classification on news (Figure 2c) and parliamentary speeches (Figure 2d).

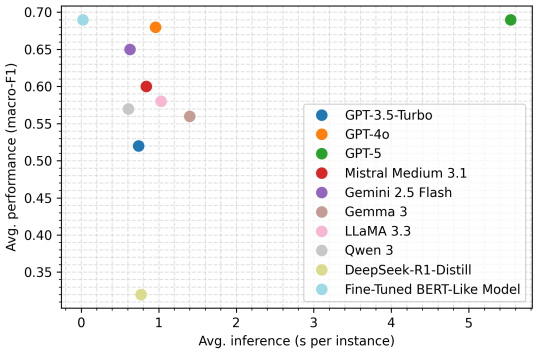


Figure 3: Comparison of models on the parliamentary topic classification based on their inference speed (seconds per instance) and performance (macro-F1 scores), both averaged across all four languages.

instance. In contrast, most LLMs have inference times between 0.6 and 1.4 seconds per instance, making them three to seven times slower for annotating the same dataset. The slowest model, GPT-5, takes 5.5 seconds per instance, which renders it impractical for large-scale automatic annotation of text collections. In this regard, fine-tuned BERT-like models offer a key advantage due to their lower computational cost and higher inference speed. Moreover, they can be trained on training data that is annotated by LLMs using the recently introduced LLM teacher-student paradigm (Kuzman and Ljubešić, 2025), which considerably reduces the effort needed to develop task-specific models.

Another limitation of LLMs, as revealed by the experiments, is their occasional deviation from the defined label set. This issue was especially noticeable in topic classification and, to a lesser extent, in genre identification. The highest rate of label hallucination was found in the DeepSeek-R1-Distill model, which produced non-existing labels for 8% of instances in the news topic test datasets and 4% in the genre test dataset. Similar issues were also observed, though much less frequently (less than 1%), with the LLaMA 3.3, Gemma 3, Qwen 3 and Mistral Medium 3.1 models. In contrast, fine-tuned

BERT-like models do not suffer from this issue, as they output probabilities for the predefined classes.

Model	Difference (sentiment)	Difference (topic)
GPT-5	0.02	0.05
GPT-4o	0.04	0.08
Gemini 2.5 Flash	0.04	0.08
Gemma 3	0.05	0.07
LLaMA 3.3	0.05	0.07
Mistral Medium 3.1	0.07	0.09
Qwen 3	0.07	0.10
GPT-3.5-Turbo	0.07	0.03

Table 3: Difference between model performance in macro-F1 scores obtained on sentiment and topic classification in parliamentary texts on English versus the average macro-F1 scores on South Slavic languages.

5.4. Performance on English versus on South Slavic languages

The sentiment identification ParlaSent and the topic classification ParlaCAP benchmark families comprise test datasets in South Slavic languages and English that were constructed with the same methodology. Thus, they also allow for a comparison of the performance of the LLMs on English, a highly resourced language, with South Slavic languages, which are significantly less represented in the pretraining and instruction-tuning datasets used to develop large language models.

As shown in Table 3, the differences in macro-F1 scores between English and the average of macro-F1 scores for South Slavic languages are relatively small for sentiment identification, ranging from 2 to 7 points. For topic classification, the performance gap is slightly larger, ranging from 3 to 10 points. This is likely due to the increased difficulty of the task, which involves greater label granularity: 22 labels compared to just 3 in sentiment classification. These findings partially confirm hypothesis H2, which stated that LLMs, when used in a zero-shot setup, perform comparably on text classification tasks in South Slavic languages as they do on English.

Interestingly, even the open-weight LLaMA 3.3 model – reported to support only eight languages, excluding the South Slavic group – does not show a substantial performance drop when applied to South Slavic languages compared to English.

6. Conclusion

In this paper, we evaluated how well current machine learning technologies handle text classification tasks in South Slavic languages. We compared fine-tuned BERT-like models with decoder-only generative large language models (LLMs) that are used in a zero-shot prompting setup across three tasks and three text domains: sentiment classification in parliamentary texts, news topic classification, topic classification in parliamentary texts, and automatic genre identification on web texts.

Our results show that LLMs used with prompting, where only a brief task description and labels were provided, achieved strong results across all tasks and languages, particularly the closed-source GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025) and Gemini 2.5 Flash (Comanici et al., 2025) models. The performance of LLMs is comparable or higher than that of fine-tuned BERT-like models specialized for the tasks. On the sentiment identification task, most open-weight and closed-source LLMs outperformed the fine-tuned model, demonstrating strong general knowledge of the notion of sentiment. For genre and topic classification, however, fine-tuning BERT-like models remains beneficial, as these tasks rely on predefined label sets and fine-tuning aligns the models more closely with the task requirements.

Interestingly, LLMs perform similarly in English and South Slavic languages, with rather minor drops in micro- and macro-F1 scores, namely a drop of 2 to 7 points in terms of macro-F1 scores on sentiment classification, and a slightly higher drop from 3 to 10 points on topic classification in parliamentary texts. This suggests that the gap in multilingual performance is smaller than expected, even for open-weight models not explicitly dedicated to these languages.

Although large language models offer impressive zero-shot performance and reduce the need for annotated data, they come with higher computational costs and are more prone to producing invalid labels. Moreover, their inference speed is at least three times slower than that of the fine-tuned BERT-like models. Thus, their use in use cases with extensive data to be processed, such as automatic enrichment of large corpora with text categories, remains impractical due to their high computational demands. In contrast, fine-tuned BERT-like models are more computationally efficient and can be better tailored to the specific characteristics of a task and its domain. They remain a practical and reliable choice for text classification tasks, especially when computational resources are limited, high inference speed is desired or output reliability is critical. Moreover, it is possible to combine the strengths of both approaches, as proposed by

the LLM teacher-student paradigm (Kuzman and Ljubešić, 2025): LLMs can be used to automatically annotate training data, reducing the need for costly and time-consuming manual annotation, while fine-tuned BERT-like models can then be trained on these datasets.

This study represents only an initial step to systematically benchmark text classification performance in South Slavic languages. Although our evaluation includes four diverse benchmark families, some of the test datasets remain relatively small. Future work will aim to increase dataset sizes, include more South Slavic languages and dialects, and introduce additional classification tasks. As new large language models continue to emerge rapidly, it will also be important to establish ongoing evaluations to track whether their performance continues to improve, particularly on South Slavic languages. Importantly, this study only evaluated the performance of LLMs in a zero-shot prompting setup. In future work, we plan to extend the evaluation to include few-shot prompting and fine-tuning on training data. To support further research and facilitate reproducibility, we have made all code, evaluation scripts, and results publicly available.[5] Additionally, we have developed an interactive dashboard[6] that enables users to explore the results of our evaluation of large language models on the tasks presented in this paper, as well as on additional commonsense reasoning tasks. The dashboard is an ongoing project that monitors the performance of newly released large language models on South Slavic languages and dialects. In future work, we plan to expand both the range of tasks and the coverage of South Slavic languages and dialects included in the dashboard.

7. Ethical Considerations and Limitations

Our study has several limitations that should be acknowledged. First, while we aimed to include a broad set of South Slavic languages, some – most notably Bulgarian – were not covered in our experiments. We assume that the performance on Bulgarian would be similar to that observed for Macedonian, given their close linguistic proximity, or the results for Bulgarian could be slightly better, as Macedonian is comparatively more low-resourced. Moreover, due to the high computational cost of evaluating the LLMs on all the test datasets and the financial cost associated with the use of closed-source models, each model was evaluated on each dataset only once. This setup prevents us from fully estimating the variance of the results, however, based on our preliminary experiments, we expect this variance to be relatively small. Finally, the scope of our evaluation remains limited in terms of test datasets, language coverage and tasks. Expanding the range of benchmarks would allow for a more comprehensive validation of our findings, particularly regarding the hypothesis that LLMs can perform on par with fine-tuned BERT-like models across diverse natural language understanding tasks, languages and language varieties.

8. Acknowledgements

We would like to thank the developers of the llm.ijs.si service (Marić et al., 2025) for establishing the LLM inference platform deployed at the Jožef Stefan Institute, which provided convenient access to the open-weight large language models used in this study. We also thank the annotators of the test datasets for their diligence and the time devoted to manual annotation, which resulted in the high-quality evaluation datasets used in this work. Lastly, we would like to thank the CLASSLA knowledge centre for South Slavic languages and the Slovenian CLARIN.SI infrastructure for their valuable support.

This work was supported in part by the projects "Spoken Language Resources and Speech Technologies for the Slovenian Language" (Grant J7-4642), "Large Language Models for Digital Humanities" (Grant GC-0002), the research programme "Language Resources and Technologies for Slovene" (Grant P6-0411), all funded by the ARIS Slovenian Research and Innovation Agency, and the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070), co-funded by the Kliping d.o.o. agency. The authors acknowledge the OSCARS project – and its ParlaCAP cascading grant project –, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This research is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

[5] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

[6] See the CLASSLA LLM Evaluation Dashboard for South Slavic Languages at https://www.clarin.si/classla-llm-dashboard/.

Marta Bañón, Miquel Esplà-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation*, pages 301–302.

Frank R Baumgartner, Christian Breunig, and Emiliano Grossman. 2019. *Comparative Policy Agendas: Theory, Tools, Data*. Oxford University Press.

Shaun Bevan. 2019. Gone Fishing. *Comparative Policy Agendas: Theory, Tools, Data*, pages 17–34.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Orphée De Clercq, Luna De Bruyne, and Véronique Hoste. 2020. News topic classification as a first step towards diverse news recommendation. *Computational Linguistics in the Netherlands Journal*, 10:37–55.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2025. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

IPTC. 2022. Groups of NewsCodes. https://iptc.org/standards/newscodes/groups/#descrncd. Accessed October 29, 2024.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Lancaster University.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage Publications.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey. *Language Resources and Evaluation*, pages 1–34.

Taja Kuzman and Nikola Ljubešić. 2025. LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, 13:35621–35633.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In *Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Taja Kuzman Pungeršek, Peter Rupnik, Daniela Širinić, and Nikola Ljubešić. 2026. Supercharging Agenda Setting Research: The ParlaCAP Dataset of 28 European Parliaments and a Scalable Multilingual LLM-Based Classification. *arXiv preprint arXiv:2602.16516*.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024a. DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Ivan Vulić, Fabian Schmidt, and Goran Glavaš. 2024b. JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 209–219.

Nikola Marić, Boshko Koloski, Damjan Demšar, Jan Jona Javoršek, and Sašo Džeroski. 2025. Running large language models locally: design and operational insights with llm.ijs.si. *International conference AI for science 2025: Ljubljana, Slovenia, 22.09.2025-26.09.2025*, page 77.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Meta. 2024. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: June 26, 2025.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv preprint arXiv:2004.03705*.

Mistral AI. 2025. Medium is the new large. https://mistral.ai/news/mistral-medium-3. Accessed: October 10, 2025.

Michal Mochtak, Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2025. Parlasent: mapping sentiment in political discourse with large language models. *Political Research Exchange*, 7(1):2508377.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036.

OpenAI. 2023. ChatGPT General FAQ. https://help.openai.com/en/articles/6783457-chatgpt-general-faq. Accessed: June 26, 2025.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed September 11, 2024.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: October 10, 2025.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Alina Petukhova and Nuno Fachada. 2023. MN-DS: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5):74.

Qwen Team. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen Team. 2024b. Qwen2.5: A party of foundation models. Accessed: June 26, 2025.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. 2025. Pushing the limit of LLM capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528.

Hang Zhao, Qile P Chen, Yijing Barry Zhang, and Gang Yang. 2024. Advancing Single and Multi-task Text Classification through Large Language Model Fine-tuning. *arXiv preprint arXiv:2412.08587*.

10. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and others. 2024. *Multilingual comparable corpora of parliamentary debates ParlaMint 4.1*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1912.

Kuzman, Taja and Ljubešić, Nikola. 2023. *Multilingual text genre classification model X-GENRE*. Hugging Face. PID https://doi.org/10.57967/hf/0927.

Kuzman, Taja and Ljubešić, Nikola. 2024a. *English-Slovenian text genre dataset X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1960.

Kuzman, Taja and Ljubešić, Nikola. 2024b. *Genre-enriched web corpora MaCoCu-Genre*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1969.

Kuzman, Taja and Ljubešić, Nikola. 2024c. *Multilingual IPTC Media Topic dataset EMMediaTopic 1.0*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1991.

Kuzman, Taja and Ljubešić, Nikola. 2024d. *Multilingual text genre classification model X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1961.

Kuzman, Taja and Ljubešić, Nikola. 2025. *Multilingual IPTC News Topic Classifier*. Hugging Face. PID https://doi.org/10.57967/hf/4709.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2025. *Multilingual ParlaCAP model for CAP Topic Classification in Parliamentary Speeches*. Hugging Face. PID https://doi.org/10.57967/hf/6684.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2026. *Multilingual training dataset for CAP policy topic classification ParlaCAP-train*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2093.

Ljubešić, Nikola and Rupnik, Peter and van Noord, Rik. 2023. *Multilingual parliamentary model XLM-R-parla*. Hugging Face. PID https://doi.org/10.57967/hf/6717.

Mochtak, Michal and Rupnik, Peter and Meden, Katja and Ljubešić, Nikola. 2023. *The multilingual sentiment dataset of parliamentary debates ParlaSent 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1868.

Rupnik, Peter and Ljubešić, Nikola and Mochtak, Michal. 2023. *Multilingual parliament sentiment regression model XLM-R-ParlaSent*. Hugging Face. PID https://doi.org/10.57967/hf/6718.

A. Appendix

A.1. Benchmarking Datasets

In this section, we provide additional information on the datasets used for benchmarking the models on sentiment identification, topic classification, and genre identification tasks in this study.

ParlaSent test datasets for sentiment classification in parliamentary speeches include Croatian, Serbian, Bosnian, and English data from the multilingual sentiment dataset of parliamentary debates ParlaSent 1.0 (Mochtak et al., 2024, Mochtak et al., 2023).[7] The dataset comprises sentences that were randomly sampled from Croatian, Serbian, Bosnian and British parliamentary corpora and manually annotated with reported inter-annotator agreement ranging from 0.53 to 0.66 in Krippendorff's alpha (Krippendorff, 2018). The annotation involved a more granular six-level sentiment polarity scale that has been mapped to a three-level sentiment polarity scale which we use in our experiments: negative (0), neutral (1), and positive (2).

GINCO datasets for automatic genre identification comprise the English EN-GINCO dataset (Kuzman et al., 2023) and a multilingual X-GINCO dataset from the AGILE benchmark for Automatic Genre Identification.[8] The test instances were sampled from the enTenTen20 English web corpus (Jakubíček et al., 2013) and the MaCoCu multilingual web corpus collection (Bañón et al., 2022). They were manually annotated by experts with a background in linguistics and computational linguistics who had experience with previous genre annotation campaigns (Kuzman et al., 2022, 2023) where they reached an acceptable inter-annotator agreement of 0.71 in nominal Krippendorff's alpha (Krippendorff, 2018). While the X-GINCO dataset comprises numerous European languages, for the purposes of this study, we focus on three South Slavic languages: Croatian, Macedonian, and Slovenian. The test datasets use the X-GENRE annotation schema (Kuzman et al., 2023) that includes the following genre labels: *Information/Explanation*, *News*, *Instruction*, *Opinion/Argumentation*, *Forum*, *Prose/Lyrical*, *Legal* and *Promotion*. While EN-GINCO and X-GINCO datasets have been annotated by the same annotator with the same schema, one should note that there are important differences between them in terms of their construction – the English test dataset was sampled randomly from the web corpus, resulting in an unbalanced label distribution, while the X-GINCO datasets were curated with more deliberate interventions to ensure a balanced label distribution and a more controlled sampling process. Consequently, the X-GINCO datasets comprise fewer ambiguous instances and could be regarded as an easier test dataset.

IPTC News Topic test datasets (Kuzman and Ljubešić, 2025) comprise Croatian and Slovenian news articles extracted from the MaCoCu-Genre web corpus collection (Kuzman and Ljubešić, 2024b) and manually annotated by one annotator. The reliability of the annotator was confirmed on a sample of data that was annotated by an additional annotator. The two annotators reached an acceptable inter-annotator agreement of 0.73 in nominal Krippendorff's alpha (Krippendorff, 2018). Text instances are annotated with 17 topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema, developed by the International Press Telecommunications Council (IPTC) (IPTC, 2022). The datasets are more or less balanced by labels.

ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) comprise parliamentary speeches in Bosnian, Croatian, English, and Serbian, sourced from the ParlaMint 4.1 dataset (Erjavec et al., 2024; Erjavec et al., 2025). These speeches were annotated by a single expert annotator using the 21 CAP categories from the official CAP schema (Baumgartner et al., 2019), along with an additional *Other* label. The datasets are approximately balanced across labels. To assess the annotation quality, the Croatian dataset was independently annotated by two additional annotators. Inter-annotator agreement between the expert annotator and the others ranged from 0.62 to 0.68 in Krippendorff's alpha, which is around the threshold of 0.67 typically considered acceptable for annotation reliability (Krippendorff, 2018).

A.2. Models

In the following subsections, we outline the models included in the evaluation – the fine-tuned BERT-like classifiers (Section A.2.1) and the open-weight and closed-source LLMs (Section A.2.2).

A.2.1. Fine-Tuned BERT-like Models

BERT (bidirectional encoder representations from transformers) deep neural models (Kenton and Toutanova, 2019) have revolutionized the field of

[7] Available in the CLARIN.SI repository at http://hdl.handle.net/11356/1585 and in the Hugging Face repository at https://huggingface.co/datasets/classla/ParlaSent.

[8] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

natural language processing (NLP), outperforming non-neural methods across various NLP tasks. They have a more complex and computationally expensive architecture featuring transformers – neural networks with self-attention mechanisms (Vaswani et al., 2017) – that significantly improves the efficiency of training the models on massive text data. Similarly to decoder-only transformer models, BERT models are pretrained on massive amounts of texts, possibly in multiple languages, which establishes their ability to encode the words and texts in high-dimensional vector spaces (Minaee et al., 2020) and enables their application even across languages in a zero-shot classification scenario. To develop BERT-based classifiers, the pretrained models are trained, that is, fine-tuned, on a training dataset comprising text instances annotated with labels. In our study, we evaluate openly-accessible multilingual fine-tuned BERT-like models that have already been developed in recent related research. Namely, we evaluate the following models:

- **IPTC News Topic classifier**[9] (Kuzman and Ljubešić, 2025) is a multilingual fine-tuned BERT-like model for news topic classification according to the top-level IPTC NewsCodes schema (IPTC, 2022). The model is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on 15,000 training text instances from the EMMediaTopic[10] dataset (Kuzman and Ljubešić, 2024c). The training dataset contains news article instances in four languages: Catalan, Croatian, Greek, and Slovenian. The training dataset was annotated using an LLM that was shown to achieve annotation reliability comparable to that of human annotators (Kuzman and Ljubešić, 2025). This approach is based on the novel methodology that uses the LLM teacher-student pipeline to develop BERT-like classifiers in the absence of manually-annotated training data.
- **XLM-R-ParlaSent** (Rupnik et al., 2023; Mochtak et al., 2024) is a domain-specific multilingual transformer model for sentiment identification in parliamentary texts. It is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on 1.72 billion words from parliamentary proceedings in 30 European languages. To develop the XLM-R-ParlaSent model,[11] the pretrained XLM-R-Parla model was fine-tuned on the ParlaSent sentiment training dataset (Mochtak et al., 2024; Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian). The training dataset[12] comprises 13,000 instances sampled from parliamentary proceedings and manually annotated with sentiment labels.
- **ParlaCAP classifier**[13] (Kuzman Pungeršek and Ljubešić, 2025; Kuzman Pungeršek et al., 2026) is a domain-specific multilingual transformer model for topic classification in parliamentary texts based on the CAP schema (Baumgartner et al., 2019). As the XLM-R-ParlaSent model, this model is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023; Mochtak et al., 2024). The XLM-R-parla model was then fine-tuned on the ParlaCAP-train dataset[14] (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026). The training dataset comprises around 30 thousand speeches from parliamentary debates from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The training dataset was annotated with the CAP categories by a GPT-4o (OpenAI, 2024) model used in a zero-shot prompting setup, following the LLM teacher-student framework (Kuzman and Ljubešić, 2025). Based on the inter-annotator agreement, calculated on a sample that was annotated by three human annotators and the LLM annotator, the agreement between the LLM and the human annotators was comparable to the agreement between the human annotators themselves. This indicates that the LLM performs as reliably as human annotators on this task, supporting its use for annotating the training data.
- **X-GENRE classifier** (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d) is a multilingual fine-tuned BERT-like model for automatic genre identification.[15] The model is based on

[9]The IPTC News Topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier.

[10]The EMMediaTopic training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/1991.

[11]The XLM-R-ParlaSent model is accessible in the Hugging Face repository at https://huggingface.co/classla/xlm-r-parlasent.

[12]The ParlaSent training and test datasets are freely available in the CLARIN.SI repository at http://hdl.handle.net/11356/1868.

[13]The ParlaCAP topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/ParlaCAP-Topic-Classifier.

[14]The ParlaCAP-train training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/2093.

[15]The X-GENRE classifier is freely available in the

the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a), which contains 1,772 text instances in Slovenian and English, manually-annotated with genre labels from the X-GENRE schema (Kuzman et al., 2023).

A.2.2. Instruction-Tuned Large Language Models

As the BERT models, decoder-only large language models are based on a transformer deep neural architecture and are pretrained on massive text collections. However, while the development of fine-tuned BERT-like classifiers necessitates large amounts of annotated training data, recent advances in the field have shown that the instruction-tuned LLMs are capable of text classification in a zero-shot or few-shot prompting setups which do not require any training data. We assess the performance of the following large language models:

- **OpenAI models**, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`) (OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`) (OpenAI, 2025). These closed-source instruction-tuned LLMs were developed by OpenAI. OpenAI states that the models are trained on large multilingual web corpora, however, specific details about the training data, procedures, and architecture are not publicly known.
- **Gemini 2.5 Flash model** (Comanici et al., 2025) is a closed-source multilingual and multimodal instruction-tuned LLM by Google DeepMind. The model is reported to be pretrained on over 400 languages (Comanici et al., 2025), however, details on the language coverage are not available.
- **Mistral Medium 3.1 model** (`mistral-medium-2508`) (Mistral AI, 2025) is a closed-source multimodal instruction-tuned model by Mistral AI. Available details on the model architecture and language coverage are very limited.
- **LLaMA 3.3 model**[16] (Meta, 2024) is an open-weight instruction-tuned multilingual LLM, developed by Meta, with 70 billion parameters. The model was pretrained on a web text collection in various languages, however, it is reported to support only 8 languages, namely, English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.
- **Gemma 3 model**[17] (Gemma Team et al., 2025) is an open-weight multilingual instruction-tuned LLM, developed by Google DeepMind. The model was pretrained on multimodal data with large quantities of multilingual texts and is reported to support over 140 languages. We use the model in 27 billion parameter size.
- **DeepSeek-R1-Distill**[18] (Guo et al., 2025) is an open-weight reasoning LLM, developed by DeepSeek AI. We use the distilled model in 14 billion parameter size, namely the `DeepSeek-R1-Distill-Qwen-14B` model. The model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that was fine-tuned using a dataset curated with the DeepSeek-R1 reasoning model. The Qwen 2.5 model provides multilingual support for over 29 languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, and Arabic.
- **Qwen 3**[19] (`Qwen3-2504`) (Yang et al., 2025) is an open-weight LLM, developed by Alibaba Cloud. We use the model with the 32 billion parameter size, namely, the `qwen3:32b` model. The model is said to support over 100 languages and dialects (Yang et al., 2025).

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). We use the quantized versions of the models as they are available through the Ollama library.[20] OpenAI models are used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[21] that provides a unified API access to various closed-source models.

To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. The same prompts were used for all open-weight and closed-source models. In Figure 4, we provide prompts that were provided to the LLMs for zero-shot text classification, namely for sentiment classification (Figure 4a), automatic genre identification (Figure 4b), news topic classification (Figure 4c) and topic classification in parliamentary

Hugging Face repository at https://doi.org/10.57967/hf/0927 and the CLARIN.SI repository at http://hdl.handle.net/11356/1961.

[16] https://ollama.com/library/llama3.3

[17] https://ollama.com/library/gemma3

[18] https://ollama.com/library/deepseek-r1:14b

[19] https://ollama.com/library/qwen3

[20] https://ollama.com/library

[21] https://openrouter.ai/

speeches (Figure 4d). For more details on the setups used to apply fine-tuned BERT-like models and instruction-tuned LLMs to the test datasets, refer to the code published on GitHub.[22]

[22] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

```
### Task
Your task is to classify the provided parliamentary text into a sentiment label, meaning that you need to recognize whether the speaker's sentiment towards the topic is negative, neutral, positive or somewhere in between. You will be provided with an excerpt from a parliamentary speech in {lang} language, delimited by single quotation marks. Always provide a label, even if you are not sure.

### Output format
Return a valid JSON dictionary with the following key: 'sentiment' and a value should be an integer which represents one of the labels according to the following dictionary: {sentiment_description}.

Text: '{text}'
```

(a) Sentiment classification.

```
### Task
Your task is to classify the following text according to genre. Genres are text types, defined by the function of the text, author's purpose and form of the text. Always provide a label, even if you are not sure.

### Output format
Return a valid JSON dictionary with the following key: 'genre' and a value should be an integer which represents one of the labels according to the following dictionary: {labels_dict}.

Text: '{text}'
```

(b) Automatic genre identification.

```
### Task
Your task is to classify the provided text into a topic label, meaning that you need to recognize what is the topic of the text. You will be provided with a news text, delimited by single quotation marks. Always provide a label, even if you are not sure.

### Output format
Return a valid JSON dictionary with the following key: 'topic' and a value should be an integer which represents one of the labels according to the following dictionary: {label_dict_with_description}.

Text: '{text}'
```

(c) News topic classification.

```
### Task
Your task is to classify the provided text into a policy agenda topic label, meaning that you need to recognize what is the predominant topic of the text. You will be provided with an excerpt from a parliamentary speech from the {parl} parliament in {language} language, delimited by single quotation marks. Always provide a label, even if you are not sure.

Follow the following rule: if the speech mentions a policy area and a policy instrument (e.g., taxes, laws), pick the label based on the area, not the instrument (e.g., annotate mortgage tax changes with 14 (Housing), law on education with 6 (Education)).

### Output format
Return a valid JSON dictionary with the following key: 'topic' and a value should be an integer which represents one of the labels according to the following dictionary: {majortopics_description}.

Text: '{text}
```

(d) Parliamentary topic classification.

Figure 4: The prompts that are provided to the LLMs for the sentiment identification task (Figure 4a), automatic genre identification (Figure 4b), and topic classification on news (Figure 4c) and parliamentary speeches (Figure 4d). The prompts comprise the description of the task and labels with short descriptions.

4.8 Final Remarks

In this chapter, we further confirmed that the X-GENRE dataset, presented in Section 3.2.2, is well-suited for training and evaluating genre classifiers. To determine which machine learning approaches provide the most robust generalization across datasets and languages, we experimented with a range of text classification methods, from traditional non-neural machine learning models to cutting-edge approaches based on large language models. To support this evaluation, we introduced two newly developed multilingual test datasets that enable model assessment in cross-lingual and cross-dataset scenarios.

The creation of new test datasets in English (EN-GINCO) and ten additional European languages (X-GINCO), presented in Section 4.1.1, fulfills the second part of Goal *G3*, which aimed to develop separate test datasets for a range of European languages. The resulting datasets cover eleven languages from three major language families – Indo-European, Turkic, and Afro-Asiatic – and include different writing systems, enabling robust evaluation of genre classifiers across both languages and scripts. Building on these datasets, we establish a publicly available benchmark, the AGILE Benchmark (Automatic Genre Identification Benchmark) on GitHub,[3] fulfilling Goal *G4*: "Establish a benchmark for cross-dataset and cross-lingual assessment of machine learning approaches in automatic genre identification, based on newly developed test datasets in English, Slovenian, and various other European languages."

By evaluating model performance in both the in-dataset scenario (Section 4.2) and the cross-dataset scenario (Section 4.3), we show that the XLM-RoBERTa model, fine-tuned on the X-GENRE training dataset, sets the state of the art for this task. On the English-Slovenian X-GENRE test dataset, it meets the threshold for reliable performance set at 0.80 for micro-F1 and macro-F1 scores. Notably, it is the only model to achieve minimally acceptable performance (micro-F1 and macro-F1 scores above 0.70) on the Slovenian portion of the test set, with a macro-F1 score of 0.76. All other evaluated models fall short by 30 or more points on the Slovenian language. These results provide empirical support for Hypothesis *H2* "*Fine-tuned deep neural Transformer-based language models outperform other machine learning methods in the task of automatic genre identification in the Slovenian language.*"

In the cross-dataset scenario on the English EN-GINCO dataset (Section 4.3), the performance of the XLM-RoBERTa model fine-tuned on the X-GENRE dataset declines. However, it remains the only evaluated model that approaches the threshold for minimally acceptable performance, achieving a micro-F1 score of 0.68 and a macro-F1 score of 0.69. This drop in performance can also be attributed to the more challenging nature of the EN-GINCO test set, as later discussed in Section 4.5.

We then further extend the evaluation to the multilingual X-GINCO datasets and show that the model achieves reliable performance on nine out of ten languages. Excluding Maltese, it reaches macro-F1 scores ranging from 0.80 to 0.95, which is consistent with Hypothesis *H3*: "*A reliable zero-shot cross-lingual performance in automatic genre identification across numerous European languages is possible by leveraging the cross-lingual capabilities of massively multilingual pre-trained Transformer models.*" The highest macro-F1 scores are observed for Ukrainian, Slovenian, and Macedonian, each exceeding 0.90. The lower performance on Maltese is likely due to the absence of this language from the XLM-RoBERTa pretraining corpus, a limitation previously noted in other NLP tasks (Chau et al., 2020; de Vries et al., 2022; Ebrahimi & Kann, 2021; Muller et al., 2021). In contrast, the model performs remarkably well on test sets using non-Latin scripts and on languages that are not closely related to the fine-tuning languages. These results suggest

[3] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

that the classifier is well-suited for genre annotation tasks across all languages included in the XLM-RoBERTa pretraining data.

These results confirm the ability of the fine-tuned XLM-RoBERTa model to generalize across both datasets and languages, which we regard as robustness. Motivated by these positive results, we release the model publicly under the name *X-GENRE classifier* on the CLARIN.SI repository[4] and the Hugging Face repository[5]. With this, we fulfill Goal *G5*: "Develop a robust multilingual genre classifier capable of generalizing across datasets and European languages by exploring various machine learning approaches, including deep neural language models."

Additionally, we experiment with a recent approach to text classification: using decoder-only instruction-tuned Transformer-based large language models (LLMs) guided by instructions in the form of prompts. This approach represents a form of zero-shot classification, as these models were not specifically trained on a genre dataset. The results in Section 4.5 show that both open-source and closed-source LLMs achieve strong zero-shot performance across all languages except Maltese. Most models exceed the estimated minimum threshold for acceptable performance, defined as a macro-F1 score of approximately 0.70. The top-performing LLMs achieve average macro-F1 scores above 0.76, with peak scores reaching up to 0.87 and surpassing the high reliability threshold of 0.80, set in Section 1.3. This level of performance is particularly impressive given that the models were not fine-tuned on the task but were only provided with a brief task description and label definitions. These findings highlight the potential of zero-shot prompting as a method for automatic genre identification. Despite this, the XLM-RoBERTa model fine-tuned on the X-GENRE training dataset continues to outperform LLMs in most languages. We argue that fine-tuned BERT-like models remain the more suitable choice for automatic genre identification, especially in large-scale annotation scenarios, as they offer higher output reliability, faster inference speed, and lower computational cost.

Finally, the results in Section 4.5 reveal clear differences in performance across large language models. For example, the gap between the GPT-3.5-Turbo model (2023) and the GPT-5 model (2025) highlights the rapid pace of progress in LLM development. The results also show that LLM performance varies significantly across languages, with open-source models outperforming closed-source ones in some cases. These findings demonstrate that the AGILE benchmark is valuable not only for assessing the state of the art in automatic genre identification, but also as a broader resource for evaluating and comparing large language models across languages. Consequently, we also integrate the EN-GINCO and X-GINCO test datasets into a broader benchmark that includes a range of text classification and commonsense reasoning tasks for South Slavic languages (Kuzman Pungeršek, Rupnik, Porupski, et al., 2026).[6] Based on this benchmark, we develop and publish an interactive dashboard "CLASSLA LLM Evaluation Dashboard for South Slavic Languages",[7] that enables interactive exploration and analysis of LLM evaluation results across languages and tasks.

[4]http://hdl.handle.net/11356/1961
[5]https://doi.org/10.57967/hf/0927
[6]https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic/
[7]https://www.clarin.si/classla-llm-dashboard/

Chapter 5

Development of a Genre Classifier Without Manually-Annotated Data

As shown in Chapter 4, recent large language models can achieve reliable performance on automatic genre identification using only a task description and genre label definitions. However, as discussed in Section 4.5, this approach is not practical for large-scale applications, such as annotating web corpora with millions of texts. In such cases, fine-tuned BERT-like models are more suitable, offering greater control over inputs, faster inference, and lower computational costs. Although closed-source LLMs deliver the best performance, they have notable drawbacks: high usage costs, lack of transparency regarding their architecture and training data, and limited reproducibility due to API-based access. In contrast, fine-tuned BERT-like models offer more flexibility and transparency. Nevertheless, a key limitation of traditional text classification methods, which include fine-tuned BERT-like models, is their reliance on high-quality task-specific training data. Creating such training datasets is challenging, time-consuming, and requires substantial human effort, as detailed in Chapter 3.

In this chapter, we propose a method that combines the strengths of traditional text classification and the recently introduced approach of zero-shot LLM prompting to mitigate the limitations of both approaches. Specifically, we introduce a text classification framework that provides an alternative to extensive and time-consuming manual annotation of training data. The proposed methodology, called the LLM Teacher-Student Framework, uses a large language model as a data annotator – the teacher model – to automatically annotate training datasets with accuracy comparable to human-level annotation. A BERT-like model – the student – is then fine-tuned on this automatically-annotated data to produce a classifier that is more computationally efficient and scalable, while achieving performance comparable to its teacher. This chapter is based on the paper *"LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification"* by Kuzman and Ljubešić (2025), enclosed in Section 5.3.

As shown in the aforementioned paper by Kuzman and Ljubešić (2025), the LLM Teacher-Student Framework enables the development of a classifier that provides high-quality annotations at minimal cost for news topic classification. The topic classification task is focused on assigning thematic labels to texts based on their content. In this respect, it is similar to automatic genre identification, since both tasks involve classifying texts into predefined categories. However, there are important differences between the two. Topic classification mainly depends on surface-level lexical information, whereas genre classification cannot rely only on keywords, but instead requires the identification of higher-level patterns, such as stylistic, structural, and syntactic features.

The framework employs an instruction-tuned decoder-only large language model (LLM) as the teacher model to develop a news topic training dataset through automatic annotation of 20,000 news articles in various languages. The reliability of the LLM as a data annotator is confirmed based on a comparison of its performance with two human annotators. The findings show that the inter-annotator agreement between each of the human annotators and the large language model is on par with the agreement between the two human annotators. The pair-wise inter-annotator agreement between the three of them ranges between 0.69 and 0.75 in Krippendorff's alpha (Krippendorff, 2018), which surpasses the threshold for acceptable agreement. Then, smaller BERT-like student models are fine-tuned on the LLM-annotated dataset. When evaluated on manually-annotated test sets, the student models achieve performance that is comparable to the teacher model, confirming that the LLM-annotated data is of sufficient quality to enable the development of fine-tuned BERT-like models that provide high performance.

In the remainder of this chapter, we apply the LLM Teacher-Student Framework methodology to the development of fine-tuned genre classifiers to demonstrate its potential for automatic genre identification. To transfer this methodology to a different task, two conditions must be met: (1) the LLM must achieve high classification performance in a zero-shot prompting scenario where it is provided only with a task description and label definitions, and (2) its agreement with human annotators must be comparable to the inter-annotator agreement between the human annotators, supporting its reliability as a data annotator. The first condition has already been met, as shown in Section 4.5, where the top-performing LLMs exceed the thresholds for minimally acceptable performance of 0.70 in macro-F1 on genre test datasets, and in some cases even achieve reliable performance beyond 0.80 in macro-F1. The second condition is evaluated in Section 5.1. Finally, to assess whether the methodology can produce reliable genre classifiers, we use an LLM to annotate the X-GENRE training dataset and compare the performance of a BERT-like model fine-tuned on the LLM-annotated data with the same model fine-tuned on the same text instances, but annotated by expert human annotators (see Section 5.2).

5.1 LLM Performance as a Data Annotator

To assess whether an LLM can substitute human annotators for automatic genre identification, we apply it to two datasets that were manually annotated by two expert annotators. We then compare inter-annotator agreement between the human annotators and between each annotator and the LLM. Specifically, we use the GPT-5 decoder-only instruction-tuned large language model (OpenAI, 2025) – the best-performing LLM from Section 4.5 – on the English EN-GINCO test set (see Section 4.1.1) and the evaluation and test splits of the Slovenian GINCO dataset (presented in Section 3.1.3). Both datasets were annotated by the same highly trained experts (see Section 3.1.2). Consistent with the experiments in Chapter 4, we map the GINCO labels to the X-GENRE schema used for fine-tuning the student model in the next step. The GPT-5 model is prompted using the same setup as in Section 4.5; specifically, it is instructed to assign an X-GENRE label to each text based on the provided task description and label definitions.

Table 5.1 presents the inter-annotator agreement between two human annotators and between each human annotator and the LLM on the GINCO and EN-GINCO datasets, using the X-GENRE schema. The agreement between the highly trained human annotators is higher, with nominal Krippendorff's alpha scores of 0.77 and 0.78. The performance of the LLM annotator lags slightly behind. However, it is important to note that the human annotators underwent extensive training and participated in numerous discussions throughout the annotation campaign to achieve this level of consistency. According to

Table 5.1: Inter-annotator agreement in terms of Krippendorff's nominal alpha between the human annotators (Ann1 and Ann2) and the LLM annotator on the Slovenian GINCO and English EN-GINCO genre datasets.

Pair	GINCO	EN-GINCO
Ann1 and Ann2	0.78	0.77
Ann1 and LLM	0.71	0.73
Ann2 and LLM	0.71	0.71

Krippendorff (2018), a nominal Krippendorff's alpha of 0.67 is considered the minimum threshold for acceptable inter-annotator agreement. The LLM annotator surpasses this threshold in both datasets, achieving notably high agreement with the human annotators, ranging from 0.71 to 0.73 in terms of Krippendorff's alpha. Moreover, the agreement scores between the LLM annotator and human annotators are significantly higher than human inter-annotator agreement reported in related genre annotation efforts (Egbert et al., 2015; Sharoff, 2018) (see Section 2.2). Based on these results, we conclude that the LLM is a viable data annotator, meeting the second requirement for applying the LLM Teacher-Student Framework to a new task.

5.2 Comparison of Models Fine-Tuned on Human-Annotated versus LLM-Annotated Training Data

To evaluate whether the proposed LLM Teacher-Student Framework can produce robust genre classifiers, we use the GPT-5 model (OpenAI, 2025) – the teacher model – to annotate the text instances in the X-GENRE training dataset with genre labels using a zero-shot prompting approach. Annotating approximately 1,800 instances using the OpenAI API batch option took only 20 minutes and cost $4, clearly demonstrating the speed and cost-efficiency of LLM-based annotation compared to manual human annotation.

We then fine-tune a base-sized XLM-RoBERTa model (Conneau et al., 2020) – the student model – on the LLM-annotated data. This setup follows the methodology described in the paper by Kuzman and Ljubešić (2025) enclosed in Section 5.3. The resulting student model is evaluated against the X-GENRE classifier (introduced in Section 3.2.2 and evaluated in Chapter 4) that was fine-tuned using the same base-sized XLM-RoBERTa model, hyperparameters, and training dataset. The only difference between the two models is that the first was fine-tuned on labels provided by an LLM, while the latter was fine-tuned on labels produced by humans in manual annotation campaigns.

Table 5.2 compares the performance of the student model – an XLM-RoBERTa model fine-tuned on the LLM-annotated X-GENRE training dataset – with its teacher model (GPT-5), used in a zero-shot prompting setup. It also compares the student model with the same base model fine-tuned on human-annotated versions of the same instances, which we refer to as the X-GENRE classifier. The results, measured in macro-F1 scores, show that the student model matches the performance of the teacher model on most test datasets and even exceeds it by 5 or more points on four of them. The student model surpasses the 0.70 macro-F1 threshold for acceptable performance on all datasets except for Maltese and the X-GENRE test set. Moreover, it exceeds the 0.80 threshold for reliable performance on half of the test datasets. These findings support two conclusions. First, the strong performance of the student model demonstrates that the LLM Teacher-Student Framework is a viable method for developing reliable genre classifiers. Second, the results show that the approach leverages the strengths of both LLMs and BERT-like models: LLMs can

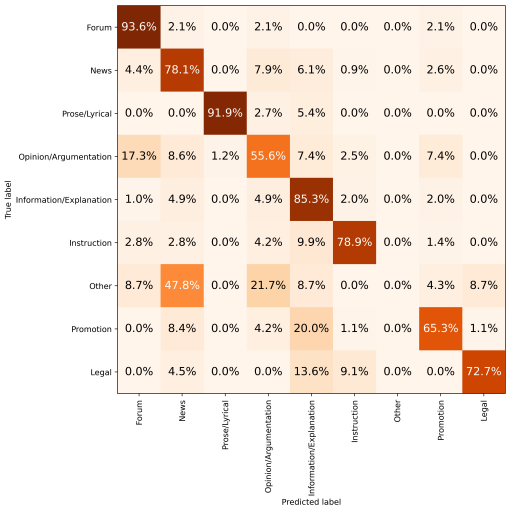
Table 5.2: Performance in terms of macro-F1 scores of the GPT-5 (teacher model), the XLM-RoBERTa model fine-tuned on the LLM-annotated X-GENRE training dataset (student model), and the XLM-RoBERTa model fine-tuned on the human-annotated X-GENRE training dataset. Models are evaluated on the test split of the X-GENRE dataset, the English EN-GINCO test dataset, and the multilingual X-GINCO test dataset. Instances where the student model matches or exceeds the performance of the teacher model are indicated with bold text.

Test Datasets	GPT-5	XLM-R & LLM Annotations	XLM-R & Human Annotations
EN-GINCO	0.72	**0.72**	0.69
X-GENRE (test)	0.68	0.67	0.79
X-GINCO: Albanian	0.75	**0.83**	0.85
X-GINCO: Catalan	0.77	**0.80**	0.84
X-GINCO: Croatian	0.72	**0.76**	0.88
X-GINCO: Greek	0.79	0.76	0.84
X-GINCO: Icelandic	0.87	0.77	0.80
X-GINCO: Macedonian	0.72	**0.80**	0.92
X-GINCO: Maltese	0.54	0.26	0.49
X-GINCO: Slovenian	0.79	**0.86**	0.94
X-GINCO: Turkish	0.85	**0.85**	0.90
X-GINCO: Ukrainian	0.83	**0.88**	0.95

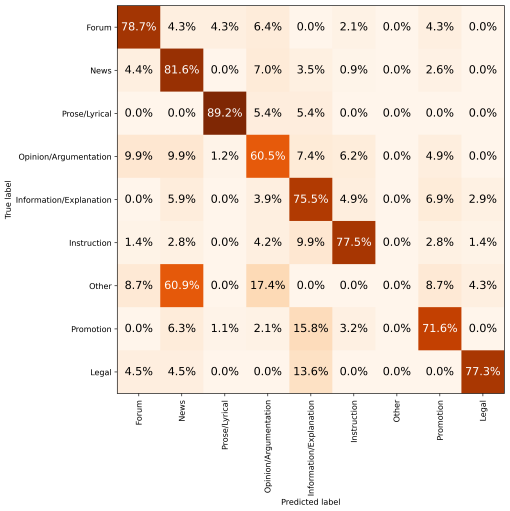
provide high-quality annotations via zero-shot prompting, while BERT-like models benefit from domain adaptation through fine-tuning on specific instances, which exposes them to the particular characteristics of the training texts. This domain-specific fine-tuning allows the student model to even outperform its teacher in some cases.

The performance of the teacher and student models on the Maltese test dataset in Table 5.2 and the label-level performance analysis in Figure 5.1 on the X-GENRE test dataset highlight an important limitation of the Teacher-Student approach. Specifically, the performance of the teacher model on individual labels or languages sets an approximate upper bound on what the student model can achieve for those labels or languages. If the teacher model fails to recognize a label accurately or does not perform well on a specific language in general, this weakness is likely to be transferred to the student model through fine-tuning on its annotations. This limitation is particularly evident for the label *Other* in Figure 5.1. Since this category includes all instances that do not align well with specific labels, it is vague by default. Human annotators rely on their judgment to assign texts to *Other*, but the prompt provided to the teacher model only loosely defines it as a category for texts that do not belong elsewhere. As a result, the LLM tends to avoid using this label and instead prefers assigning more concrete categories. Figure 5.1a shows that the teacher model did not assign any instance to this category in the X-GENRE test dataset. This omission is then carried over to the student model, which fails to recognize or predict this label, as shown in Figure 5.1b.

The low performance on the *Other* category also helps explain why the student model performs worse on the EN-GINCO and X-GENRE test datasets compared to the X-GINCO dataset, which does not include this label. When we exclude instances labeled as *Other* from the EN-GINCO and X-GENRE test datasets, both the teacher and student models show significantly improved performance, as demonstrated in Table 5.3. Without this label, both models reach reliable performance, exceeding 0.80 in macro-F1 and micro-F1



(a) Teacher model performance.



(b) Student model performance.

Figure 5.1: Label-level performance of the teacher GPT-5 model (Figure 5.1a) and the student XLM-RoBERTa model (Figure 5.1b) on the X-GENRE test split. Confusion matrices are normalized by true labels, showing for each actual class the percentage of samples assigned to each predicted class.

on the EN-GINCO dataset, and even outperform the X-GENRE classifier. Similarly, their performance on the X-GENRE test dataset improves by 10 points. The *Other* category is less critical than other labels for the end use cases, as ambiguous cases could be also filtered out of the annotated text collection using confidence scores derived from the outputs of the fine-tuned XLM-RoBERTa model. The results in Table 5.3 thus even further confirm the applicability of the LLM Teacher-Student approach and highlight the strong performance of the teacher and student models.

Table 5.3: Performance in terms of micro-F1 and macro-F1 scores of the teacher model, the student model and the X-GENRE model, evaluated on the X-GENRE and EN-GINCO test datasets from which the label *Other* had been removed.

Dataset	Model	Macro-F1	Micro-F1
EN-GINCO	GPT-5	0.84	0.84
	XLM-R & LLM Annotations	0.84	0.81
	XLM-R & Human Annotations	0.75	0.71
X-GENRE (test)	GPT-5	0.78	0.76
	XLM-R & LLM Annotations	0.77	0.75
	XLM-R & Human Annotations	0.82	0.81

To conclude, the results presented in this section demonstrate that replacing human annotators with a large language model for training data annotation is a viable strategy for developing robust and efficient genre classifiers. The XLM-RoBERTa model fine-tuned on LLM-annotated data achieves strong performance across all test datasets except Maltese. However, it is important to note that the student model rarely matches the performance of the same base model fine-tuned on human-annotated data. This highlights the added value of human annotation, which, despite being more time-consuming, labor-intensive,

and costly, tends to produce higher-quality training data. If the disadvantages of the manual annotation campaigns are not a major limitation for a specific use case, we thus still recommend using human annotators. We do not argue that human annotation will become obsolete. Especially for evaluation purposes, human-annotated test sets remain essential for ensuring reliable results.

In scenarios where large-scale annotation of multilingual training data is required, human annotation may not be feasible due to resource constraints. In such cases, LLM-based annotation – provided that the conditions put forward earlier in this chapter are met – offers a highly effective alternative. For example, we could improve the multilingual genre classifier by extracting tens of thousands of instances from the MaCoCu web corpora (Bañón, Esplà-Gomis, Forcada, García-Romero, et al., 2022) in ten languages and annotating them with genre labels using an LLM. This would allow us to efficiently expand the X-GENRE training dataset and potentially improve model performance across the ten languages in the X-GINCO test dataset. Moreover, the LLM Teacher-Student Framework is particularly useful for handling infrequent labels. In cases where a label appears too rarely in a manually-annotated dataset, thousands of candidate examples can first be annotated automatically using the LLM based on label definitions. Human annotators can then validate these pre-annotated examples of the infrequent label, making it feasible to incorporate rare categories into the training without the need for very large manually-annotated datasets.

Finally, the LLM Teacher-Student methodology is particularly valuable for languages that lack manually-annotated training data suitable for the task. If the LLM used has strong natural language understanding capabilities in the target language and has demonstrated good performance on the task in other languages, it can be leveraged for the development of high-quality training data. This can then improve the performance of the fine-tuned BERT-like model on that language. This approach could be especially useful for Maltese, where none of the models in our experiments achieved high performance in genre classification. As LLMs continue to evolve and expand their capabilities, including better coverage of low-resource languages, we expect that future models, especially those with broader multilingual pretraining, will be better equipped to handle languages like Maltese.

5.3 Related Paper: LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 January 2025, accepted 17 February 2025, date of publication 24 February 2025, date of current version 28 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3544814

RESEARCH ARTICLE

LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification

TAJA KUZMAN[1,2] **AND NIKOLA LJUBEŠIĆ**[1,3]
[1]Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
[3]Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia

Corresponding author: Taja Kuzman (taja.kuzman@ijs.si)

This work was supported in part by the Projects "Embeddings-Based Techniques for Media Monitoring Applications" co-funded by the Kliping d.o.o. Agency and the Slovenian Research and Innovation Agency (ARIS) under grant L2-50070, in part by the "Large Language Models for Digital Humanities" funded by ARIS under Grant GC-0002, and in part by the Research Programme "Language Resources and Technologies for Slovene" funded by ARIS under Grant P6-0411.

ABSTRACT With the ever-increasing number of news stories available online, classifying them by topic, regardless of the language they are written in, has become crucial for enhancing readers' access to relevant content. To address this challenge, we propose a teacher-student framework based on large language models (LLMs) for developing multilingual news topic classification models of reasonable size with no need for manual data annotation. The framework employs a Generative Pretrained Transformer (GPT) model as the teacher model to develop a news topic training dataset through automatic annotation of 20,000 news articles in Slovenian, Croatian, Greek, and Catalan. Articles are classified into 17 main categories from the Media Topic schema, developed by the International Press Telecommunications Council (IPTC). The teacher model exhibits high zero-shot performance in all four languages. Its agreement with human annotators is comparable to that between the human annotators themselves. To mitigate the computational limitations associated with the requirement of processing millions of texts daily, smaller BERT-like student models are fine-tuned on the GPT-annotated dataset. These student models achieve high performance comparable to the teacher model. Furthermore, we explore the impact of the training data size on the performance of the student models and investigate their monolingual, multilingual, and zero-shot cross-lingual capabilities. The findings indicate that student models can achieve high performance with a relatively small number of training instances, and demonstrate strong zero-shot cross-lingual abilities. Finally, we publish the best-performing news topic classifier, enabling multilingual classification with the top-level categories of the IPTC Media Topic schema.

INDEX TERMS Multilingual text classification, IPTC, large language models, LLMs, news topic, topic classification, training data preparation, data annotation.

I. INTRODUCTION

Topic classification is a significant asset in the news industry, enabling automatic identification of news topics and facilitating readers' access to content that aligns with their interests. However, developing a robust news topic classifier is challenging, particularly because of the absence of manually annotated data, which are especially scarce for non-English languages. Manual annotation, although invaluable, is an expensive and labor-intensive process that cannot be quickly extended to numerous languages.

Despite the emergence of zero-shot approaches that employ Generative Pretrained Transformers (GPTs) that have demonstrated impressive results in various natural language

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

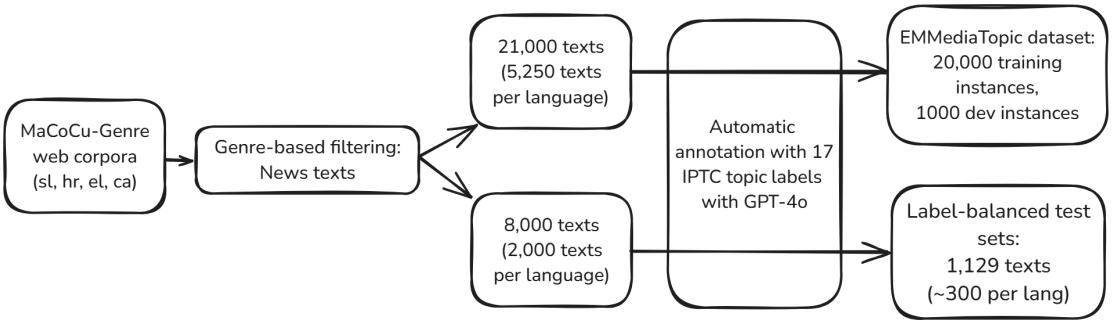


FIGURE 1. Pipeline for the preparation of training and test datasets for the classification of news topics.

processing tasks in diverse languages [1], [2], [3], their application in settings with extensive data to be processed remains impractical due to their high computational demands. A more suitable technology for this setting is a smaller BERT-like language model; however, such models require thousands of annotated instances for effective fine-tuning for a certain task.

In this paper, we introduce an approach to developing annotated training data and multilingual BERT-like news topic classifiers through a teacher-student framework based on large language models (LLMs). Our methodology leverages the power of GPTs that we use as teacher models to automatically annotate news articles with top-level Media Topic labels, introduced by the International Press Telecommunications Council (IPTC) [4]. This automatically annotated dataset is then used to train a smaller XLM-RoBERTa-based student model, addressing the computational demands of the news media industry, where topic annotation needs to be scaled to millions of data points.

In our experiments, we train and evaluate the models on multilingual training and test datasets comprising four diverse languages: Catalan, Croatian, Greek, and Slovenian. The selection of these languages for evaluation is based on their availability in the MaCoCu corpora collection [5] to ensure a high level of comparability between language-specific datasets. Furthermore, these languages exhibit varying degrees of linguistic relatedness, with Croatian and Slovenian being closely related, and the other two languages belonging to separate branches of the Indo-European language family. This distinction provides valuable insights for cross-lingual experiments.

In this work, our aim is to address the following research questions, pertaining to the task of IPTC news topic classification:

(RQ1) Can the teacher LLM, used as a data annotator for news topic classification, achieve annotation quality comparable to that of human annotators?
(RQ2) How much data, annotated by the teacher model, are required for the student model to achieve performance comparable to the teacher model?
(RQ3) Is it necessary to include the target language in the training data, or does the fine-tuned student model exhibit satisfactory zero-shot cross-lingual capabilities?
(RQ4) Does fine-tuning a student model on target-language monolingual data yield better outcomes than training on a multilingual dataset of equal size?

In summary, this study makes the following contributions. We investigate the feasibility of the LLM teacher-student framework for the development of accurate and computationally efficient multilingual news topic classifiers for languages that do not have readily accessible manually annotated training data suitable for the task. Specifically, we focus on single-label multi-class text classification employing top-level categories of the IPTC NewsCodes Media Topic schema. This study assesses the applicability of a GPT model for data annotation by comparing the level of agreement between the model and human annotators to the level of agreement among the human annotators themselves. Furthermore, we examine the classification performance of the student model relative to its teacher model, while also exploring the impact of training data size on the student model's performance, as well as its zero-shot cross-lingual and multilingual capabilities. Motivated by these positive results, we publish the best performing model in the Hugging Face repository (https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier), providing an IPTC news topic classification model that can be applied to any of the 100 languages covered by the XLM-RoBERTa model. To the best of our knowledge, this is the first openly-available multilingual topic classifier that uses the (top-level) IPTC Media Topic categories.

The remainder of this paper is organized as follows. Section II provides an overview of the existing literature on IPTC news topic classification and the application of GPT models to data annotation. In Section III, we introduce our methodology for automatic annotation of training data by leveraging the zero-shot capabilities of large language models. In Section IV, we describe the manual annotation campaign that was undertaken to develop a test set for

evaluating the performance of both teacher and student models, and we evaluate the performance of the GPT model as a data annotator compared to human annotators. Section V presents the experiments that involve fine-tuning BERT-like models on varying sizes of training data, and the analysis of their monolingual, multilingual, and zero-shot cross-lingual performance. Section VI concludes the paper with a discussion of the main findings and suggestions for future research. Additionally, the Appendix provides more details on the annotation guidelines and the description of the top-level IPTC Media Topic labels (Section A), and the prompt used for automatic data annotation (Section B).

II. RELATED WORK

The International Press Telecommunications Council (IPTC) is a global organization dedicated to the development and promotion of industry standards for the exchange of news data. Among its notable contributions is the maintenance of IPTC NewsCodes controlled vocabularies that are widely adopted by major media providers, including Agence France-Presse, Associated Press, and Reuters [6]. The NewsCodes vocabularies include the Media Topic vocabulary, which has been established as a global standard to ensure consistent coding of news metadata across diverse news providers [7]. The Media Topic taxonomy is structured hierarchically, with 17 topic labels at the top level. The entire taxonomy encompasses more than 1,000 topic labels, organized across up to five levels of granularity [8].

Multiple previous studies investigated news topic classification [9], [10], [11], [12], [13]. However, there is no manually annotated reference dataset that can be used directly for the training and evaluation of IPTC Media Topic classifiers. The existing topic datasets have several limitations:

1) **Inconsistent Topic Schemata:** Many datasets use their own topic schemata [9], [12], [13], rendering the results of their experiments highly dataset dependent and consequently incomparable with other studies.
2) **Lack of Manual Annotation:** Many topic datasets did not undergo manual annotation. Instead, topic categories were derived from source media websites [12], [13] or automatically annotated using topic modeling [9] based on the word2vec algorithm [14].
3) **Use of a Deprecated IPTC Schema:** Some datasets use the outdated IPTC Subject Codes schema (see, for instance, [11] and [15]), which was replaced by the IPTC Media Topic schema in 2010. Additionally, the IPTC Media Topic schema itself is subject to frequent updates, occurring every two to three months [16], which poses additional challenges for the development of a stable reference dataset and classifiers.

Recently, two news topic datasets have been introduced, both annotated with the IPTC Media Topic schema: the MN-DS dataset [17], comprising approximately 10,000 English news articles, and the EventDNA dataset [18], containing around 1,800 Dutch news articles. Despite their potential, both datasets have faced criticism regarding the reliability of manually annotated labels, as highlighted in related studies that employed these datasets for machine learning experiments [19], [20]. Beyond problems with reliability, these datasets exhibit additional limitations. The MN-DS dataset was derived from the NELA-GT-2018 dataset [21], which was originally developed for misinformation research. As a result, it includes sources known to disseminate fake news, which may compromise the performance of models trained on it when applied to mainstream news texts. The EventDNA dataset, on the other hand, is limited by its annotations being performed at the lowest levels of the IPTC Media Topic taxonomy. Specific sublabels do not always align well with top-level labels, which makes the dataset less useful for automatic text classification on top-level labels. Additionally, while EventDNA is freely accessible, restrictions on third-party sharing do not allow experiments with closed-source large language models.

Prior research has investigated a range of machine learning models for the automatic classification of texts into IPTC news topics. The experiments encompass both traditional non-neural models, such as logistic regression, the Naive Bayes classifier, and support vector machines [17], [20], as well as deep learning techniques, particularly BERT-like Transformer models [11], [17], [19]. Most of these studies have demonstrated that Transformer-based models consistently achieve state-of-the-art performance, indicating their potential in the domain of news topic classification [17], [19].

Recently introduced Generative Pretrained Transformer (GPT) models, commonly known as large language models (LLMs), have demonstrated impressive performance across a range of text classification tasks, even when used in a zero-shot prompting fashion that does not require any training data [1], [22], [23]. Furthermore, recent research in the field of natural language processing (NLP) has embraced "LLMs as catalysts for redefining the landscape of data annotation in machine learning and NLP" [24]. Specifically, researchers are exploring the potential of GPT-based annotation of training data to replace time-consuming manual annotation, with the aim of using these annotations to fine-tune or improve the performance of other GPT or BERT-like models [24], [25], [26], [27]. Although LLMs have been employed to generate and annotate instruction-tuning datasets [28], [29], [30], their use for the annotation of text classification data has been less explored. Meng et al. [31] contributed to this area by employing a GPT model to generate both texts and labels for training datasets through a zero-shot prompting approach, and by fine-tuning a smaller language model on the generated training dataset. In contrast to this approach, our study refrains from generating synthetic data to mitigate potential LLM biases. Instead, we use authentic news articles sourced from the web and employ a GPT model exclusively for classifying these texts into Media Topic labels. This GPT-based data annotation approach was also shown to be promising in the domain of sentiment analysis, where student

models fine-tuned on datasets annotated by a GPT model exhibited comparable performance to models fine-tuned on datasets annotated by human annotators [27].

A recent investigation into the applicability of GPT models for IPTC news topic classification was conducted by Fatemi et al. [20] who evaluated a GPT model on the MN-DS dataset [17]. The applicability of GPT models for topic annotation was evaluated by training various machine learning models on the GPT-annotated dataset and comparing their performance against models trained on the same dataset, but manually annotated. The findings indicate the promising performance of the GPT model used as a data annotator. However, it is important to note that the performance of the model was not directly evaluated on a manually annotated test set. Consequently, the accuracy of the GPT model in this task remains undetermined.

Building on existing research, our study explores the use of GPT-based data annotation to fine-tune a BERT-like topic classifier. This research addresses multiple gaps in the literature. First, we assess the performance of both the teacher GPT model and fine-tuned student models on reliably manually annotated data. Second, we compare the annotation reliability of the GPT model with the reliability of human annotators. Additionally, our research involves machine learning experiments on multiple languages, which offer valuable insights into the models' multilingual and cross-lingual performance.

III. LLM-BASED DATASET DEVELOPMENT

In this section, we present our methodology for the automatic annotation of training data that is used for fine-tuning a multilingual Transformer model which employs top-level IPTC Media Topic labels. The proposed methodology leverages three resources that have recently become available:

- **MaCoCu Corpora** [5]: The MaCoCu corpora collection comprises comparable web corpora in 10 less-resourced European languages, including Slovenian, Croatian, Catalan, Greek, Turkish, Icelandic, Ukrainian, and others. Developed through web crawling, MaCoCu corpora represent some of the largest text collections available for these languages. They encompass high-quality texts [32] from a wide range of sources and text genres [33].
- **X-GENRE Classifier** [1]: This multilingual Transformer-based genre classifier categorizes texts in various languages into a set of genre labels, such as *News*, *Promotion*, and *Opinion/Argumentation*. The X-GENRE classifier allows us to efficiently extract relevant news content from general web corpora, which is crucial for the present task that is focused on news articles.
- **Multilingual GPT Models**: The multilingual Generative Pretrained Transformer (GPT) models have demonstrated impressive zero-shot classification capabilities across numerous languages and tasks. Inter alia, these models have shown high performance on South Slavic languages, including Croatian and Slovenian, which are included in our experiments [34]. Specifically, we employ the GPT-4o model (version `gpt-4o-2024-05-13`) [35] provided by the OpenAI API, which outperformed other open- and closed-source GPT models in our preliminary experiments.

A. EXTRACTION OF NEWS TEXTS AND PRE-PROCESSING

The pipeline for developing training and test data for news topic classification using the LLM teacher-student framework is illustrated in Fig. 1. The datasets used in this study are sourced from the Catalan (ca) [36], Croatian (hr) [37], Greek (el) [38], and Slovenian (sl) [39] MaCoCu web corpora that have been annotated with genres (genre-annotated datasets are available as part of the MaCoCu-Genre corpus collection [40]). Given the focus of IPTC topic classification on news articles, we extract from the web corpora the subset of news articles, identified with the X-GENRE classifier [1]. The X-GENRE classifier provides highly reliable classification into the *News* genre, as well as eight other genre categories. Evaluation on a multilingual, manually annotated test set, sampled from the MaCoCu corpora, revealed F1 scores for the *News* label ranging from 0.82 in Catalan to 0.95 in Croatian [41]. Furthermore, the dataset undergoes additional pre-processing, wherein the texts are truncated to the initial 512 words, which is a constraint imposed by the BERT-like models.

B. AUTOMATIC ANNOTATION

Automatic annotation involves using the GPT-4o model to assign 17 top-level IPTC Media Topic labels to news texts. We use labels from the Media Topic schema version from October 24, 2023. The annotation is applied to two separate samples, randomly extracted from the pre-processed news text collection described in the previous section: a first batch of 21,000 texts is annotated for the development of training and development datasets, and a second batch of 8,000 texts is annotated to create a test set. Both batches comprise equal numbers of instances in the four languages, namely Catalan (ca), Croatian (hr), Greek (el), and Slovenian (sl). The GPT-4o model is used in a zero-shot prompting manner. It is instructed to assign to each text one of the 17 main IPTC Media Topic labels, such as *politics* and *society*. The prompt (see Appendix B) was constructed based on the results of the preliminary experiments. It includes the task description and a list of labels with their descriptions (provided in Appendix A). The annotation of 29,000 texts cost approximately 230€ and took approximately six hours. This cost and time investment are deemed minimal compared to the resources that would be required for the manual annotation of a similar volume of instances.

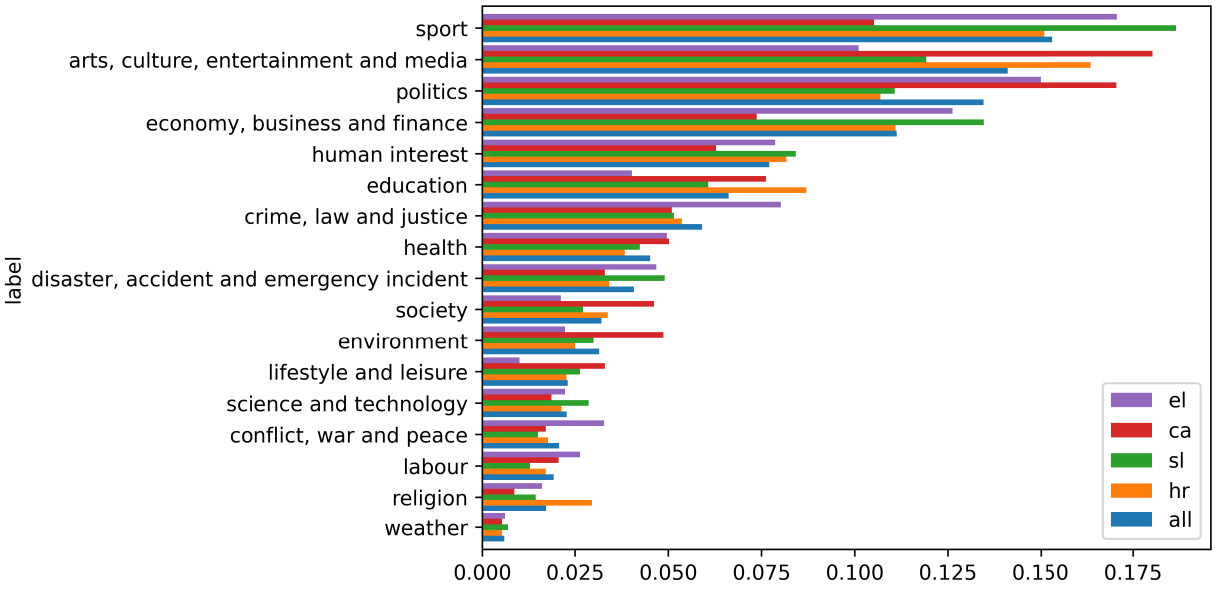


FIGURE 2. Distribution of labels in the GPT-annotated EMMediaTopic dataset comprising training and development splits (in percentages).

C. TRAINING AND DEVELOPMENT DATASETS

To construct the training and development datasets, the first annotated batch of 21,000 instances is split into subsets of 20,000 and 1,000 texts, respectively, while maintaining stratification based on the GPT-assigned labels. The resulting dataset, which comprises both splits, has been published under the name "the EMMediaTopic dataset" in the CLARIN.SI repository (http://hdl.handle.net/11356/1991) [42].

The distribution of labels within the EMMediaTopic dataset is shown in Fig. 2. The dataset is shown to be relatively balanced by labels, with the most prevalent label (*sport*) accounting for approximately 15% of the instances. Moreover, the distribution of labels is consistent across the four languages. The least represented labels are *weather*, accounting for less than 1% of instances, and *religion* and *labour*, each assigned to less than 2% of the texts.

D. TEST DATASET

The test dataset is derived from the second batch of 8,000 automatically annotated texts, comprising texts in the same four languages as the training data, namely Slovenian, Croatian, Catalan, and Greek. To assess the models' performance across all 17 top IPTC Media Topic labels, the test set is balanced by GPT-assigned labels, which is achieved by selecting 18 instances per label per language (or fewer if there are fewer than 18 label instances in the batch). The selection of texts to be manually annotated comprised 1,199 instances with a balanced distribution across languages. Finally, the test set was manually annotated to obtain human gold labels. Details on the annotation campaign are provided in the next section.

E. TEACHER MODEL PERFORMANCE

The first application of the manually annotated test set was to measure the performance of the GPT-4o model that was used for automatic annotation. The model demonstrates consistently high performance, with an average macro-F1 score of 0.731 and a micro-F1 score of 0.722. Its performance across different languages is stable, with a 5-point difference between the highest macro-F1 score (on Slovenian) and the lowest (on Catalan), as shown in Table 1.

TABLE 1. GPT-4o performance on each language in the human-labeled test set in micro-F1 and macro-F1 scores, averaged across three prediction iterations, with standard deviation included.

	Micro-F1	Macro-F1
Hr	0.714 ± 0.002	0.721 ± 0.001
Ca	0.703 ± 0.002	0.702 ± 0.001
Sl	0.741 ± 0.000	0.748 ± 0.001
El	0.730 ± 0.009	0.738 ± 0.009
Entire Test Set	0.722 ± 0.002	0.731 ± 0.003

IV. MANUAL ANNOTATION OF TEST DATA

The test sample, balanced by GPT-assigned labels and the four languages, is manually annotated by a single annotator. The annotator was provided with annotation guidelines with the descriptions of the labels (see Appendix A) and received brief training on the annotation process. The description of the labels is based on the official definitions provided by the IPTC [43]. However, since these definitions offer only a high-level understanding of the distinctions between labels, we enriched the descriptions with examples of their lower-level labels to facilitate a more precise and consistent annotation.

TABLE 2. Pair-wise inter-annotator agreement in terms of the nominal Krippendorff's alpha.

Annotators	Krippendorff's alpha
1st ann & 2nd ann	0.728
1st ann & GPT-4o	0.693
2nd ann & GPT-4o	0.752

In addition to the standard 17 top-level IPTC Media Topic labels, three additional labels have been introduced to assist the annotator in identifying and excluding texts that are not suitable for the purposes of our task: *do not know* (highly challenging instance lacking a clear topic), *not news* (text of a different genre, e.g., *Promotion*), and *multiple* (presence of multiple texts in a single instance). The manual annotation process excluded 70 instances (5.83%) as unsuitable. Among these, 4% were excluded because of the absence of a clear topic, 2% because they contained multiple texts in a single instance, and 0.1% due to incorrect genre classification. The final test sample consists of 1,129 text instances. The test dataset is available on request from the corresponding author. It is not publicly released to prevent its integration into large language models (LLMs) during their pretraining or fine-tuning phase to maintain the integrity of future model evaluations.

A. INTER-ANNOTATOR AGREEMENT

To assess the reliability of human and LLM annotations, a subset of 339 instances from the test set, balanced by labels and languages, was annotated by a second annotator. Human annotators received the same guidelines, provided in Appendix A, and participated in a preliminary test annotation which was followed by a discussion to resolve the disagreements.

The inter-annotator agreement is evaluated using Krippendorff's alpha metric [44], a statistical measure frequently employed in content analysis, social sciences, and machine learning to determine the reliability of multiple human annotators in data classification. The metric accounts for the likelihood of agreement occurring by chance, with a value of 1 indicating perfect agreement and a value of 0 indicating agreement equal to chance [45]. Theoretically, reliable inter-annotator agreement is represented by alpha values of 0.8 or higher, whereas an alpha of 0.667 is considered the threshold for acceptable annotation reliability [44]. Nonetheless, this threshold value is frequently not achieved in manual annotation campaigns for text classification tasks. This shortcoming can be attributed to the high subjectivity and complexity of annotation tasks, which often involve cases where characteristics of multiple labels are present [46]. This problem has also been demonstrated in tasks similar to topic classification, such as automatic genre identification [47], sentiment annotation [48], [49], [50], and detection of hate speech [51], [52].

In our experiments, the nominal Krippendorff's alpha between the two annotators reached 0.728. This outcome is deemed satisfactory as it exceeds the threshold for acceptable reliability. Furthermore, this level of inter-annotator agreement is comparable to or exceeds the agreement observed in similar text-level classification tasks. At the same time, the moderate level of agreement highlights the challenging nature of this task, as multiple topic labels may be discussed in certain news articles.

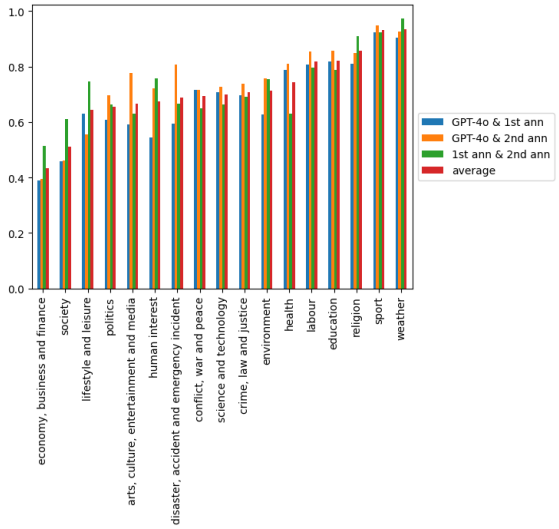


FIGURE 3. Label-level inter-annotator agreement in terms of the nominal Krippendorff's alpha.

TABLE 3. Intra-annotator agreement in terms of the nominal Krippendorff's alpha for the human annotator (1st annotator) and the LLM teacher model.

Annotator	Krippendorff's alpha
GPT-4o	0.934
Human annotator (1st ann)	0.796

As shown in Table 2, the inter-annotator agreement between each of the human annotators and the LLM model was on par with the agreement between the two human annotators. This finding addresses Research Question 1 (RQ1), demonstrating that the LLM model is as capable of data annotation for this task as human annotators.

We also assess the agreement between the human and LLM annotators at the label level. Fig. 3 illustrates the substantial variation in inter-annotator agreement across topic labels, highlighting labels that are less clearly defined, which leads to confusion among annotators. The labels *sport* and *weather* are well comprehended by both human and LLM annotators, achieving high inter-annotator agreement above 0.90 in terms of Krippendorff's alpha. Labels *religion*, *lifestyle and leisure*, and *society* present challenges for the LLM, with human annotators demonstrating significantly higher agreement. The label *economy, business and finance* is particularly problematic for both humans and the LLM, with an average inter-annotator agreement of 0.43.

B. INTRA-ANNOTATOR AGREEMENT

To further assess the reliability of the annotations provided by humans and LLMs, we instruct both a human annotator and the GPT-4o model to re-annotate the subset that was previously used to calculate the inter-annotator agreement. Annotators are given the same instructions as in the initial annotation process, and the labels previously assigned to the texts are hidden from them. Intra-annotator agreement is evaluated using Krippendorff's alpha measuring consistency between two annotation runs conducted by the same annotator. Table 3 presents the intra-annotator agreement for the human and LLM annotators. The human annotator achieved a relatively high Krippendorff's alpha of approximately 0.80, indicating a reliable level of self-agreement. The GPT-4o model demonstrated even greater consistency with Krippendorff's alpha of 0.93, highlighting the reliability of large language models in producing consistent responses. However, while such consistency is advantageous, more diverse responses, similar to those obtained from the human annotator, can be beneficial for various research directions, including learning from disagreement [53].

V. STUDENT MODEL FINE-TUNING EXPERIMENTS

In this section, we describe the experiments that involve fine-tuning of student BERT-like models on training data that were automatically annotated by the teacher GPT model. In Section V-A, we present the student model which is used in the experiments that examine (1) the impact of training data size on classification performance (Section V-B), (2) the model's label-level performance (Section V-C), and (3) the model's monolingual, multilingual, and cross-lingual capabilities (Section V-D).

A. STUDENT MODEL

As the student model, we select the large-sized XLM-RoBERTa model [54] (available on the Hugging Face repository: https://huggingface.co/FacebookAI/xlm-roberta-large), which we fine-tune on the teacher-annotated EMMediaTopic dataset. This BERT-like encoder model, pretrained on hundred languages, has demonstrated strong multilingual and cross-lingual performance across various text classification tasks, which also involved the languages used in our experiments [1], [32], [55]. The optimal hyperparameters were identified via a hyperparameter search performed on the development dataset. Appendix C provides a comprehensive description of the selected hyperparameters.

The models, fine-tuned on the teacher-annotated training dataset, are evaluated on the test dataset via micro-F1 and macro-F1 scores to assess the performance at both the instance and label levels, respectively. Each model was trained and tested at least three times to gain an understanding of the performance consistency. The reported scores represent the mean values calculated across all iterations and are accompanied by a standard deviation. In the experiments related to the training data size, five iterations were performed per data point owing to the addition of the subset selection process, which was an additional source of variation of the results.

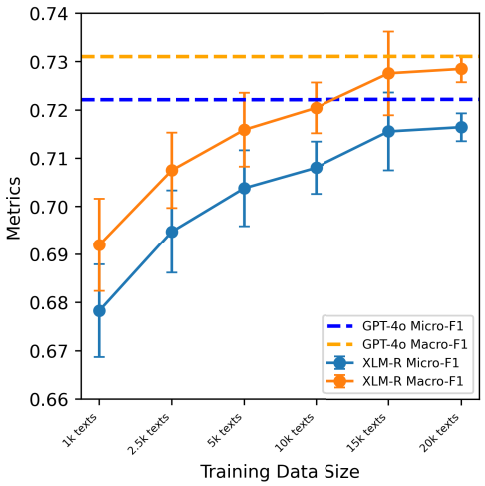


FIGURE 4. Performance in micro-F1 and macro-F1 scores of the XLM-RoBERTa (XLM-R) model fine-tuned on various sizes of training data, compared to the zero-shot GPT-4o performance as the upper limit. The scores are averaged across five iterations of fine-tuning and evaluation, each using different random sample of a specified size, drawn from the training dataset.

B. IMPACT OF THE TRAINING DATA SIZE

In this section, we address Research Question 2 (RQ2). First, we explore the feasibility of achieving high performance of BERT-like models on the news topic classification task when fine-tuned on automatically annotated data. Second, we investigate the minimum amount of training data required to attain a performance level comparable to that of the teacher model. To this end, we fine-tune the XLM-RoBERTa model on the subsets from the original training set, which are stratified by label size and balanced across languages. The smaller training datasets comprise 1,000, 2,500, 5,000, 10,000, and 15,000 instances.

Fig. 4 shows the impact of the training data size on the model performance. The performance of the GPT-4o model is considered to be an upper limit, as it served as the teacher model for annotating the training dataset for the student model. The level of improvement begins to plateau after 15,000 instances. Notably, student models trained on datasets comprising 15,000 data points or more exhibit performance levels that are comparable to those of the teacher GPT model. There is less than a one-point difference in macro-F1 and micro-F1 scores between the student and teacher models. These findings confirm the feasibility of achieving high performance in the news topic classification task using a student model trained on data annotated by a teacher model.

Fine-tuning student models on training datasets comprising 15,000 data points or more provides optimal results; however, the findings also reveal that the models exhibit

TABLE 4. Performance in macro-F1 scores of monolingual and multilingual models trained on 5,000 instances and evaluated on each language in the test split, and on the entire test split. The results of the monolingual training and evaluation are highlighted in gray color. The highest score for each language-specific test set is highlighted in bold. GPT-4o performance is added as the upper limit, as the models were trained on its predictions. The reported scores represent the average results across three iterations of training and evaluation, with the standard deviation provided.

	Hr Test	Ca Test	Sl Test	El Test	Entire Test Set
Hr Model	0.701 ± 0.017	**0.672** ± 0.005	0.733 ± 0.014	**0.739** ± 0.011	0.716 ± 0.009
Ca Model	0.706 ± 0.012	0.671 ± 0.008	0.728 ± 0.005	0.715 ± 0.014	0.710 ± 0.007
Sl Model	**0.711** ± 0.007	0.660 ± 0.016	0.736 ± 0.014	0.721 ± 0.013	0.713 ± 0.005
El Model	0.674 ± 0.004	0.662 ± 0.012	0.716 ± 0.011	0.706 ± 0.013	0.695 ± 0.006
Multilingual	0.707 ± 0.011	0.656 ± 0.024	**0.741** ± 0.007	0.729 ± 0.004	0.714 ± 0.009
GPT-4o	0.721 ± 0.001	0.702 ± 0.001	0.748 ± 0.001	0.738 ± 0.009	0.731 ± 0.003

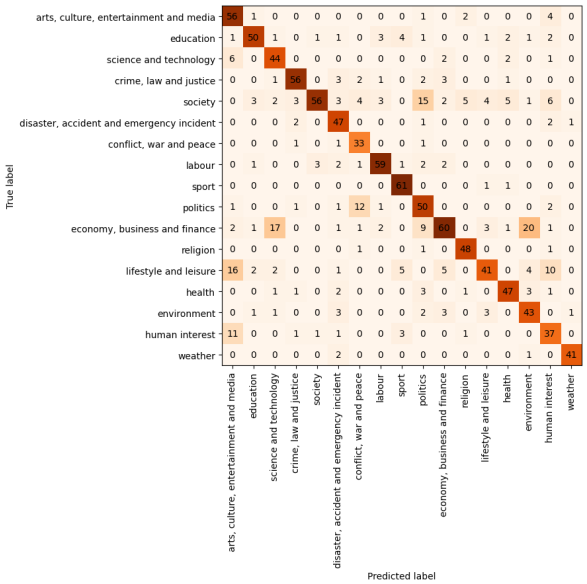


FIGURE 5. Confusion matrix for the best-performing XLM-RoBERTa student model.

substantial performance even with minimal training data. Specifically, student models trained on datasets comprising 1,000–5,000 instances achieve average micro-F1 scores ranging from 0.678 to 0.704 and macro-F1 scores between 0.692 and 0.716. Notably, the student model trained on only 1,000 instances demonstrates a performance that is only four points lower than that of the teacher GPT model.

C. LABEL-LEVEL PERFORMANCE

The best-performing version of the XLM-RoBERTa student model, namely, a version trained on 15,000 instances, has been made available in the Hugging Face repository (accessible at https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier). This model attains a macro-F1 score of 0.746 and micro-F1 score of 0.734. Further insights into the performance of the model at the label level can be obtained from the confusion matrix depicted in Fig. 5, which displays the true versus predicted labels.

The labels that are most accurately predicted by the model are *weather*, *sport*, and *religion*, with label-level F1 scores above 0.89. These findings are consistent with the label-level inter-annotator agreement shown in Fig. 3 (see Section IV-A), which identified these labels as those with the highest consensus among human annotators. Interestingly, despite the *weather* label being represented by less than 1% of the instances of the training dataset, this did not negatively impact the performance of the model. The model demonstrated a near-perfect classification accuracy for this label, achieving an F1 score of 0.94. This result confirms the usefulness of the classifier also for the labels that are less represented in the training dataset.

In contrast, the most challenging labels for the XLM-RoBERTa student model are *lifestyle and leisure*, *human interest*, and *economy, business and finance*, as evidenced by label-level F1 scores ranging from 0.59 to 0.62. This is not surprising, given that these categories have also been identified as challenging for both human and LLM annotators. As shown in Fig. 5, the label *economy, business and finance* is frequently misclassified as *science and technology* when the texts pertain to the sales of technology products; or as *environment* in instances where the texts address electricity infrastructure. The label *human interest* is often mistaken for *arts, culture, entertainment and media*, because both categories encompass content related to celebrities. The former is typically associated with tabloid-style presentations of celebrity lives, whereas the latter is intended to cover more serious topics. Similarly, the label *lifestyle and leisure* frequently overlaps with both *human interest* and *arts, culture, entertainment and media*, as it encompasses individuals' recreational activities, which may intersect with topics represented by the other two labels. These overlaps illustrate the complexities inherent in IPTC topic classification, which pose challenges to both manual annotation and automatic classification. These challenges arise primarily from the broad nature of top-level labels, which often leads to intersections among categories. A potential solution to these problems is to conduct classification with labels from the lower levels of the Media Topic schema, which are more specific but also more challenging due to their high number.

D. MONOLINGUAL, MULTILINGUAL AND CROSS-LINGUAL PERFORMANCE

The preceding section demonstrated the potential for achieving high performance using the LLM teacher-student setup for the task of news topic classification. This section explores

the applicability of student models to languages on which they were not fine-tuned. We analyze the performance of fine-tuned models in monolingual, multilingual, and cross-lingual settings, addressing Research Questions 3 (RQ3) and 4 (RQ4).

Specifically, we fine-tune the XLM-RoBERTa model on monolingual subsets extracted from the original training dataset comprising only Croatian (hr), Slovenian (sl), Catalan (ca), or Greek (el) instances. Each monolingual subset consists of 5,000 instances in the target language, and the subsets have similar label distributions. Additionally, the evaluation includes the multilingual model, trained on samples comprising 5,000 instances (as described in Section V-B), distributed equally across the four languages. The training dataset for the multilingual model is limited to 5,000 instances to ensure that the multilingual model is trained on a dataset of equivalent size to that of the monolingual models and to eliminate any potential influence of the dataset size on the results.

The models are evaluated on the manually annotated test dataset, split into target language subsets. Each monolingual test subset comprises approximately 300 instances. The reported scores are averaged across the three iterations of fine-tuning and evaluation.

Table 4 presents the performance of the monolingual and multilingual models across four languages included in the test set. In a zero-shot cross-lingual scenario – where the models are evaluated on a language different from the one they were fine-tuned on –, the models demonstrate relatively high macro-F1 scores, ranging from 0.66 to 0.74. Moreover, the zero-shot cross-lingual performance is often comparable to, and in most cases even exceeds, the models' performance in monolingual and multilingual settings – where the models are evaluated on the language which is included in their (either monolingual or multilingual) fine-tuning dataset. These findings, pertaining to Research Question 3 (RQ3), demonstrate that student models exhibit high performance even when applied to languages that are not part of the fine-tuning training dataset.

Second, we compare the performance of student models on a target language when they are fine-tuned on 5,000 instances of (1) monolingual data in the target language, or (2) multilingual data that are balanced across the target language and three additional languages. Through these experiments, we address Research Question 4 (RQ4), which focuses on investigating whether the development of language-specific models offers any advantages over the development of a multilingual model trained on an equivalent amount of data, consisting of instances in the target language and other languages. If the experiments would show that monolingual results are significantly higher, this might require to host potentially many language-specific models for cases where top performance is important enough. The results presented in Table 4, where the monolingual performance is highlighted in gray, however, reveal that the multilingual model achieves performance that is comparable to, or on certain language-specific test sets surpasses that of the model fine-tuned only on the target language. This finding highlights the benefits of training student models on multiple languages, through which the robustness of the model is quite likely to improve. More importantly, this result allows the final inference procedure to be very simple, with a single multilingual model that can be applied to any text written in one of the 100 supported languages on which the XLM-RoBERTa model was pretrained [54].

VI. CONCLUSION

This paper presents a novel methodology for developing annotated training data using a teacher-student framework based on large language models. By fine-tuning a smaller Transformer model on the data annotated by a GPT model, this methodology allows for scalable classification of news topics without the need for manual annotation of the training data. The code for the development of GPT-annotated training data and XLM-RoBERTa topic classifiers is publicly accessible (https://github.com/TajaKuzman/IPTC-Media-Topic-Classification).

The IPTC Media Topic training dataset is developed by extracting news articles from web corpora in four languages (Slovenian, Croatian, Greek, and Catalan) based on automatic genre identification using the X-GENRE classifier [1] and automatic topic annotation using the GPT model as a teacher model. The EMMediaTopic dataset [42] has been made freely available in the CLARIN.SI repository (http://hdl.handle.net/11356/1991).

The GPT model, used in a zero-shot prompting fashion, demonstrates high performance across all languages and labels, with an average macro-F1 score of 0.731 and micro-F1 score of 0.722. The inter-annotator agreement between the two human annotators is comparable to the agreement between the teacher LLM model and each of the human annotators. Furthermore, the intra-annotator agreement analysis reveals that the GPT model demonstrates a significantly higher consistency in topic labeling than the human annotator. The results indicate that large language models possess a comparable ability to human annotators in labeling training data with news topic categories, but with less variation in the output.

GPT models are computationally too demanding for scaling topic annotation to millions of data points that regularly need to be annotated, as is the case in the news media industry. Thus, we develop smaller BERT-like models, considered as student models, by fine-tuning them on the GPT-annotated EMMediaTopic training dataset. We show that the developed training data are of a sufficient quality to achieve high performance of the student models. This finding confirms the feasibility of the LLM teacher-student framework for the development of multilingual news topic classifiers. Specifically, the XLM-RoBERTa-based student models, when fine-tuned on 15,000 or more data points, demonstrate a performance comparable to that of the GPT teacher model, with a difference of less than one point in both

the micro-F1 and macro-F1 scores. Motivated by positive results, we publish the best-performing student model on Hugging Face (https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier), providing the first open-source IPTC Media Topic classification model for the top layer of the Media Topic schema that can be applied to any language covered by the XLM-RoBERTa model. The model was trained on 15,000 instances in four languages, and it achieves a micro-F1 score of 0.734 and a macro-F1 score of 0.746.

Furthermore, we delve deeper into the constraints of the performance of the student models, investigating their ability to generalize in a zero-shot cross-lingual scenario. The experiments with student models fine-tuned either on monolingual or multilingual training data and evaluated on four languages show high zero-shot cross-lingual performance, yielding macro-F1 scores between 0.66 to 0.74. Moreover, the zero-shot cross-lingual performance exceeds the models' performance in monolingual and multilingual settings. These results indicate that incorporating the target language within the training dataset is not essential for achieving a high topic classification performance in that language.

Further experiments concerning the scalability of IPTC classification across multiple languages investigate the potential advantages of developing language-specific student models compared with constructing a multilingual model fine-tuned on the same amount of data in four languages, thereby less data in the target language. An analysis of the performance of monolingual versus multilingual models across four target languages reveals that the multilingual model achieves a performance that is either comparable or superior to that of models fine-tuned exclusively on the target language. This finding suggests that multilingual models are a more effective approach to catering to multiple languages. These models provide high performance across multiple languages while requiring less annotated data in the target language and incurring smaller processing and storage costs compared to the development of individual models for each language.

In this study, we employ only the 17 top-level categories of the IPTC Media Topic hierarchical schema. The analysis of the agreement between human annotators and the LLM annotator for each category, as well as the confusion matrix for the fine-tuned model, highlighted challenges associated with the ambiguity of certain top-level labels. In future work, we plan to extend our experiments to include more fine-grained categories from the lower levels of the IPTC schema, which are more useful for some users and news media providers. We consider following the hierarchical approach proposed by [19] and [20], where the top label is assigned to the text first with the classifier presented in this paper, and then based on the predicted label, a more specialized lower-layer classifier is selected to make deeper level-two predictions. Moreover, our current experiments involve single-label classification, which does not account for the multifaceted nature of some texts in which multiple top-level topics could be useful to the end user. To address this, we plan to move towards a multi-label classification setting based on the information on the confidence of the student model.

Motivated by positive results, we intend to expand our experiments involving the LLM teacher-student framework to topic classification across various domains, such as parliamentary proceedings and web corpora. Furthermore, we aim to explore the applicability of the proposed methodology to additional text classification tasks, including automatic genre identification, which is recognized as a challenging task for human annotators [47].

Furthermore, future work could include a more in-depth exploration of potential biases that might emerge in the fine-tuned student model. The biases may be introduced from the original sources, since the training data is sourced from a large collection of online news articles, as well as from the large language model used for annotating these training data. Given that the news topic categories include sensitive subjects such as *conflict, war and peace*, *religion*, and *society*, it is important to consider the potential biases originating from both the training data and the large language model while deploying the automatic news classification system in real-world applications.

VII. LIMITATIONS

While our LLM teacher-student methodology provides promising results, it is important to acknowledge that the success of this approach is fundamentally dependent on the credibility of the teacher model, as its performance represents the upper limit of what can be achieved with the framework. In our study, the GPT model demonstrated strong performance in the news topic classification task, achieving high macro-F1 and micro-F1 scores and exhibiting annotation reliability comparable to that of human annotators. These findings highlight the feasibility of the framework for this specific task, providing high-quality annotations at minimal cost. However, this level of performance may not necessarily be replicable across other natural language processing tasks or languages, where the conventional method of training on manually annotated data may still be preferable. Therefore, it is crucial that researchers assess the teacher model on manually annotated test datasets to ensure its reliability and comparability to human annotators prior to applying the LLM teacher-student approach to their use cases.

APPENDIX A
ANNOTATION GUIDELINES

A. GENERAL INSTRUCTIONS

Annotate the provided texts with the IPTC topic labels, provided below. If it is hard to choose between two labels, you can also help yourself by searching through the taxonomy of sublabels (https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html) where you might find the specific topic of your text and see under which of out top-level labels it fits (but do not spend too much time on

TABLE 5. IPTC Media Topic labels and their descriptions, which have been constructed by the authors of this paper.

Label	Description
arts, culture, entertainment and media	News about cinema, dance, fashion, hairstyle, jewellery, festivals, literature, music, theatre, TV shows, painting, photography, woodworking, art exhibitions, libraries and museums, language, cultural heritage, news media, radio and television, social media, influencers, and disinformation.
conflict, war and peace	News about terrorism, wars, wars victims, cyber warfare, civil unrest (demonstrations, riots, rebellions), peace talks and other peace activities.
crime, law and justice	News about committed crime and illegal activities, the system of courts, law and law enforcement (e.g., judges, lawyers, trials, punishments of offenders).
disaster, accident and emergency incident	Man-made or natural events resulting in injuries, death or damage, e.g., explosions, transport accidents, famine, drowning, natural disasters, emergency planning and response.
economy, business and finance	News about companies, products and services, any kind of industries, national economy, international trading, banks, (crypto)currency, business and trade societies, economic trends and indicators (inflation, employment statistics, GDP, mortgages, ...), international economic institutions, utilities (electricity, heating, waste management, water supply).
education	All aspects of furthering knowledge, formally or informally, including news about schools, curricula, grading, remote learning, teachers and students.
environment	News about climate change, energy saving, sustainability, pollution, population growth, natural resources, forests, mountains, bodies of water, ecosystem, animals, flowers and plants.
health	News about diseases, injuries, mental health problems, health treatments, diets, vaccines, drugs, government health care, hospitals, medical staff, health insurance.
human interest	News about life and behaviour of royalty and celebrities, news about obtaining awards, ceremonies (graduation, wedding, funeral, celebration of launching something), birthdays and anniversaries, and news about silly or stupid human errors.
labour	News about employment, employment legislation, employees and employers, commuting, parental leave, volunteering, wages, social security, labour market, retirement, unemployment, unions.
lifestyle and leisure	News about hobbies, clubs and societies, games, lottery, enthusiasm about food or drinks, car/motorcycle lovers, public holidays, leisure venues (amusement parks, cafes, bars, restaurants, etc.), exercise and fitness, outdoor recreational activities (e.g., fishing, hunting), travel and tourism, mental well-being, parties, maintaining and decorating house and garden.
politics	News about local, regional, national and international exercise of power, including news about election, fundamental rights, government, non-governmental organisations, political crises, non-violent international relations, public employees, government policies.
religion	News about religions, cults, religious conflicts, relations between religion and government, churches, religious holidays and festivals, religious leaders and rituals, and religious texts.
science and technology	News about natural sciences and social sciences, mathematics, technology and engineering, scientific institutions, scientific research, scientific publications and innovation.
society	News about social interactions (e.g., networking), demographic analyses, population census, discrimination, efforts for inclusion and equity, emigration and immigration, communities of people and underrepresented groups (LGBTQ, older people, children, indigenous people, etc.), homelessness, poverty, societal problems (addictions, bullying), ethical issues (suicide, euthanasia, sexual behaviour) and social services and charity, relationships (dating, divorce, marriage), family (family planning, adoption, abortion, contraception, pregnancy, parenting).
sport	News about sports that can be executed in competitions, e.g., basketball, football, swimming, athletics, chess, dog racing, diving, golf, gymnastics, martial arts, climbing, etc.; sport achievements, sport events, sport organisation, sport venues (stadiums, gymnasiums, ...), referees, coaches, sport clubs, drug use in sport.
weather	News about weather forecasts, weather phenomena and weather warning.

```
    structured_prompt_label_description = f"""
            ### Task
            Your task is to classify the provided text into a topic label, meaning that
you need to recognize what is the topic of the text. You will be provided with a news
text, delimited by single quotation marks. Always provide a label, even if you are not
sure.

            ### Output format
            Return a valid JSON dictionary with the following key: 'topic' and a value
should be an integer which represents one of the labels according to the following
dictionary: {label_dict_with_description}.
            """
```

FIGURE 6. Prompt used for automatic annotation of data with the GPT-4o model.

it, if it is very hard to decide, rather annotate it as "do not know").

B. ADDITIONAL LABELS FOR HARD CASES

We are interested only in reasonable and informative instances. That is why we added to the list of labels three more categories for you to "discard" the text:

- *do not know*: if it is impossible to decide under which label this instance fits.
- *not news*: if the text is not a news article, for example, it is an advertisement, recipe, legal text, blog post, etc.
- *multiple*: if the instance is not a single text, but consists of multiple texts.

1) IPTC LABEL DESCRIPTION

Table 5 presents the IPTC Media Topic labels along with their descriptions, as provided to the human annotators and GPT model.

APPENDIX B
PROMPT FOR AUTOMATIC ANNOTATION

Fig. 6 shows the prompt presented to the GPT model for the purpose of automatic annotation. The prompt includes descriptions of the labels, presented in Table 5.

APPENDIX C
FINE-TUNING HYPERPARAMETERS FOR THE XLM-ROBERTA MODELS

In the fine-tuning experiments, we use the following hyperparameters that were determined through a hyperparameter search on the development dataset: a learning rate of $8e^{-6}$, a train batch size of 32, and a maximum sequence length of 512 tokens. The optimal number of training epochs was found to depend on the size of the training data, as shown in Table 6.

TABLE 6. Number of epochs used for fine-tuning the XLM-RoBERTa model, depending on the size of the training data (number of instances).

Training Data Size	Epoch Number
20,000	3
15,000	5
10,000	9
5,000	10
2,500	22
1,000	24

REFERENCES

[1] T. Kuzman, I. Mozetič, and N. Ljubešić, "Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 3, pp. 1149–1175, Sep. 2023.

[2] H. Wibowo, E. Fuadi, M. Nityasya, R. E. Prasojo, and A. Aji, "COPAL-ID: Indonesian language reasoning with local culture and nuances," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2024, pp. 1404–1422.

[3] A. Chowdhery et al., "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[4] IPTC. *Media Topics*. Accessed: Oct. 29, 2024. [Online]. Available: https://iptc.org/standards/media-topics/

[5] M. Bañón, M. Esplá-Gomis, M. L. Forcada, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, and P. Rupnik, "MaCoCu: Massive collection and curation of monolingual and bilingual data: Focus on under-resourced languages," in *Proc. 23rd Annu. Conf. Eur. Assoc. Mach. Transl.*, 2022, pp. 301–302.

[6] IPTC. *NewsCodes*. Accessed: Oct. 29, 2024. [Online]. Available: https://iptc.org/standards/newscodes/

[7] IPTC. *What is IPTC?* Accessed: Oct. 29, 2024. [Online]. Available: https://iptc.org/about-iptc/

[8] IPTC. *Groups of NewsCodes*. Accessed: Oct. 29, 2024. [Online]. Available: https://iptc.org/standards/newscodes/groups/#descrncd

[9] H. Roberts, R. Bhargava, L. Valiukas, D. Jen, M. M. Malik, C. S. Bishop, E. B. Ndulue, A. Dave, J. Clark, B. Etling, R. Faris, A. Shah, J. Rubinovitz, A. Hope, C. D'Ignazio, F. Bermejo, Y. R. Benkler, and E. Zuckerman, "Media cloud: Massive open source collection of global news on the open web," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 15, May 2021, pp. 1034–1045.

[10] A. N. Chy, Md. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," in *Proc. 16th Int. Conf. Comput. Inf. Technol.*, Mar. 2014, pp. 366–371.

[11] M. Pranjic, M. Robnik-Šikonja, and S. Pollak, "An evaluation of BERT and Doc2Vec model on the IPTC subject codes prediction dataset," in *Proc. 24th Int. Multiconference Inf. Soc., Data Mining Data Warehouses Conf.*, 2021, pp. 25–28.

[12] R. Misra, "News category dataset," 2022, *arXiv:2209.11429*.

[13] I. Kosem, J. Čibej, K. Dobrovoljc, T. Kuzman, and N. Ljubešić, "Spremljevalni korpus trendi in avtomatska kategorizacija," *Slovenščina 2.0, Empirical Appl. Interdiscipl. Res.*, vol. 11, no. 1, pp. 161–188, Sep. 2023.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[15] STT. *Finnish News Agency Archive 1992–2018, Source*. Accessed: Oct. 29, 2024. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2019041501

[16] IPTC. *NewsCodes Guidelines*. Accessed: Oct. 29, 2024. [Online]. Available: https://www.iptc.org/std/NewsCodes/guidelines/

[17] A. Petukhova and N. Fachada, "MN-DS: A multilabeled news dataset for news articles hierarchical classification," *Data*, vol. 8, no. 5, p. 74, Apr. 2023.

[18] C. Colruyt, O. De Clercq, T. Desot, and V. Hoste, "EventDNA: A dataset for Dutch news event extraction as a basis for news diversification," *Lang. Resour. Eval.*, vol. 57, no. 1, pp. 189–221, Mar. 2023.

[19] O. D. Clercq, L. D. Bruyne, and V. Hoste, "News topic classification as a first step towards diverse news recommendation," *Comput. Linguistics Netherlands J.*, vol. 10, pp. 37–55, Dec. 2020.

[20] B. Fatemi, F. Rabbi, and A. L. Opdahl, "Evaluating the effectiveness of GPT large language model for news classification in the IPTC news ontology," *IEEE Access*, vol. 11, pp. 145386–145394, 2023.

[21] J. Nørregaard, B. D. Horne, and S. Adali, "NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, Jul. 2019, pp. 630–638.

[22] N. Ljubešić, N. Galant, S. Benčina, J. Čibej, S. Milosavljević, P. Rupnik, and T. Kuzman, "DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects," in *Proc. 11th Workshop NLP Similar Lang., Varieties, Dialects*, 2024, pp. 89–98.

[23] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech," in *Proc. Companion ACM Web Conf.*, Apr. 2023, pp. 294–297.

[24] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation and synthesis: A survey," 2024, *arXiv:2402.13446*.

[25] Y. Zeng, Y. Mu, and L. Shao, "Learning reward for robot skills using large language models via self-alignment," 2024, *arXiv:2405.07162*.

[26] Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang, and C. Gan, "Principle-driven self-alignment of language models from scratch with minimal human supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 2511–2565.

[27] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing, "Is GPT-3 a good data annotator?" in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 11173–11195.

[28] Z. Wang, C.-L. Li, V. Perot, L. Le, J. Miao, Z. Zhang, C.-Y. Lee, and T. Pfister, "CodecLM: Aligning language models with tailored synthetic data," in *Proc. Findings Assoc. Comput. Linguistics*, 2024, pp. 3712–3729.

[29] C. Xu, D. Guo, N. Duan, and J. McAuley, "Baize: An open-source chat model with parameter-efficient tuning on self-chat data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 6268–6278.

[30] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 13484–13508.

[31] Y. Meng, J. Huang, Y. Zhang, and J. Han, "Generating training data with language models: Towards zero-shot language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 462–477.

[32] R. van Noord, T. Kuzman, P. Rupnik, N. Ljubešić, M. Esplà-Gomis, G. Ramírez-Sànchez, and A. Toral, "Do language models care about text quality? Evaluating web-crawled corpora across 11 languages," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval.*, 2024, pp. 5221–5234.

[33] T. Kuzman, P. Rupnik, and N. Ljubešić, "Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora," in *Proc. 10th Workshop NLP Similar Lang., Varieties Dialects*, 2023, pp. 91–103.
[34] N. Ljubešić, T. Kuzman, P. Rupnik, I. Vulić, F. Schmidt, and G. Glavaš, "JSI and WüNLP at the DIALECT-COPA shared task: In-context learning from just a few dialectal examples gets you quite far," in *Proc. 11th Workshop NLP Similar Languages, Varieties, Dialects*, 2024, pp. 209–219.
[35] OpenAI. (2024). *Hello GPT-4o*. Accessed: Sep. 11, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/
[36] M. Bañón, M. Chichirau, M. Esplá-Gomis, M. L. Forcada, A. Galiano-Jiménez, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, and J. Zaragoza-Bernabeu. (2023). *Catalan Web Corpus MaCoCu-Ca 1.0*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1837
[37] M. Bañón, M. Chichirau, M. Esplá-Gomis, M. L. Forcada, A. Galiano-Jiménez, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. Pla Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, and J. Zaragoza-Bernabeu. (2023). *Croatian Web Corpus MaCoCu-HR 2.0*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1806
[38] M. Bañón, M. Chichirau, M. Esplá-Gomis, M. L. Forcada, A. Galiano-Jiménez, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, and J. Zaragoza-Bernabeu. (2023). *Greek Web Corpus MaCoCu-EL 1.0*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1839
[39] M. Bañón, M. Chichirau, M. Esplá-Gomis, M. L. Forcada, A. Galiano-Jiménez, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. Pla Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, and J. Zaragoza-Bernabeu. (2023). *Slovene Web Corpus MaCoCu-SL 2.0*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1795
[40] T. Kuzman and N. Ljubešić. (2024). *Genre-enriched Web Corpora MaCoCu-Genre*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1969
[41] T. Kuzman and N. Ljubešić. (2023). *Multilingual Text Genre Classification Model X-GENRE*. Hugging Face. [Online]. Available: https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier
[42] T. Kuzman and N. Ljubešić. (2024). *Multilingual IPTC Media Topic Dataset EMMediaTopic 1.0*. Slovenian Language Resource Repository. [Online]. Available: http://hdl.handle.net/11356/1991
[43] IPTC. *IPTC Media Topic NewsCodes Tree View*. Accessed: Oct. 29, 2024. [Online]. Available: https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html
[44] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA, USA: Sage, 2018.
[45] K. Krippendorff, "Measuring the reliability of qualitative text analysis data," *Qual. Quantity*, vol. 38, no. 6, pp. 787–800, Dec. 2004.
[46] J.-C. Klie, R. E. D. Castilho, and I. Gurevych, "Analyzing dataset annotation quality management in the wild," *Comput. Linguistics*, vol. 50, no. 3, pp. 817–866, Sep. 2024.
[47] T. Kuzman and N. Ljubešić, "Automatic genre identification: A survey," *Lang. Resour. Eval.*, pp. 1–34, Nov. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10579-023-09695-8
[48] W. van Atteveldt, M. A. C. G. van der Velden, and M. Boukes, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Commun. Methods Measures*, vol. 15, no. 2, pp. 121–140, Apr. 2021.
[49] F. Lind, M. Gruber, and H. G. Boomgaarden, "Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs," *Commun. Methods Measures*, vol. 11, no. 3, pp. 191–209, Jul. 2017.
[50] A. Pelicon, M. Pranjić, D. Miljković, B. Škrlj, and S. Pollak, "Zero-shot learning for cross-lingual news sentiment classification," *Appl. Sci.*, vol. 10, no. 17, p. 5993, Aug. 2020.
[51] J. A. Leite, D. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for Brazilian portuguese: New dataset and multilingual analysis," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, Brazil, 2020, pp. 914–924.
[52] N. Ljubešić, D. Fišer, and T. Erjavec, "The FRENK datasets of socially unacceptable discourse in Slovene and English," in *Proc. 22nd Int. Conf.*, Ljubljana, Slovenia. Cham, Switzerland: Springer, Sep. 2019, pp. 103–114.
[53] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey," *J. Artif. Intell. Res.*, vol. 72, pp. 1385–1470, Dec. 2021.
[54] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.
[55] M. Ulčar, A. Žagar, C. S. Armendariz, A. Repar, S. Pollak, M. Purver, and M. Robnik-Šikonja, "Evaluation of contextual embeddings on less-resourced languages," 2021, *arXiv:2107.10614*.

TAJA KUZMAN received the M.S. degree in translation between Slovenian, French, and English from the University of Ljubljana, Slovenia, in 2019. She is currently pursuing the Ph.D. degree in information and communication technologies with the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.

Since 2021, she has been a Research Assistant with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana. She is currently a Co-Leader of CLASSLA, the CLARIN ERIC knowledge center, which offers expertise on language resources and technologies for South Slavic languages. She is one of the main developers of the CLASSLA-web corpora, the largest text collections for these languages, and conducts evaluations of NLP models on South Slavic languages and dialects. As the author or co-author of 68 resources, she is one of the main contributors of text datasets and language technologies for Slovenian and other European languages at the CLARIN.SI repository. Her research interests include language resource collection and curation and the development of technologies for automatic annotation using deep learning, including text classification tasks, such as automatic genre identification.

NIKOLA LJUBEŠIĆ received the Ph.D. degree from the University of Zagreb on the topic of event detection in news data streams, in 2009.

Currently, he is a Senior Researcher with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. He is also with the Faculty for Computer and Information Science, University of Ljubljana, and with the Institute for Contemporary History, Ljubljana. His research interests include computational linguistics and computational social science, with an emphasis on South Slavic linguistic and cultural space. He is the co-author of the CLASSLA-Stanza linguistic processing pipeline and CLASSLA-web corpora, both covering the range of South Slavic languages. He has co-authored the training data and various benchmarks for the same language group. His general interests in language variation has motivated him to expand his focus from textual modality to speech modality, inside which he is developing a methodology to scale spoken data collection and automatic enrichment. He is co-leading the CLASSLA knowledge center for South Slavic languages with the first author.

Chapter 6

Conclusions

This thesis addressed the complex task of automatic genre identification by proposing novel methodologies for schema creation and manual annotation, and by investigating suitable machine learning techniques. Throughout the research process, the development of the genre classifier was guided by three important criteria: (1) robustness, (2) multilingual capability, and (3) high usability for downstream applications, namely, for scalable automatic annotation of large web corpora, and for effective genre-based text selection in language technology development and corpus linguistic research.

First, our examination of existing genre schemata and datasets in Chapter 2 revealed that none can be directly applied to our study. The existing high-coverage schemata provide a useful foundation, however, they have several limitations: excessive label granularity, which negatively impacts the reliability of manual annotations; limited comparability across schemata, hindering the potential merging of multiple existing genre datasets; abstract label names that may not be comprehensible to end users; and closed label sets that do not account for instances outside predefined categories. To address these challenges, we set out to develop new genre schemata that rectify the limitations identified in previous work. Our proposed schemata, presented in Chapter 3, are based on existing schemata to allow for potential mapping between genre schemata and the integration of existing genre datasets into our machine learning experiments. This approach is particularly important given the lack of comparability between datasets in previous work, which has led to dataset-specific experiments and results that provide limited insight into the robustness of the developed genre classifier, that is, its generalization capabilities when applied to other datasets.

Until recently, research on automatic genre identification has predominantly focused on the English language, with limited availability of non-English genre datasets. This thesis addresses this gap by developing a manually-annotated genre dataset in Slovenian, which is one of the target languages in our research (see Chapter 3). To ensure that the Slovenian genre dataset accurately represents real-world web texts, we follow established best practices from prior studies, including the extraction of random instances from web corpora and consideration of the temporal diversity of resources. Having conducted a manual annotation campaign for the development of the Slovenian genre dataset, we demonstrate that achieving acceptable inter-annotator agreement in genre identification annotation campaigns is feasible. This is accomplished through a manual annotation approach that is based on a well-designed genre schema, comprehensive annotation guidelines, and the employment of expert annotators.

To improve the performance of classifiers trained on manually-annotated genre datasets, we merge the Slovenian dataset with two additional manually-annotated English datasets. This integration is based on a novel joint schema to which existing schemata are mapped

(see Section 3.2 in Chapter 3). Our findings indicate that models fine-tuned on this joint dataset achieve superior results compared to models fine-tuned on other existing datasets, even those that are up to six times larger. This outcome demonstrates that the proposed dataset is more effective for training genre classifiers and reinforces the notion that dataset quality is more important than size.

Having created a multilingual training and test genre dataset – which we have shown to be highly effective for training genre classifiers – we proceed to the next phase detailed in Chapter 4: experimenting with various text classification methodologies to determine which of them provides the most robust generalization across datasets and languages. To this end, we develop two additional test datasets encompassing eleven languages that span ten branches across three major language families and use three different scripts. This linguistic diversity allows for a comprehensive evaluation of a classifier's performance across both languages and writing systems. Our experiments, which include in-dataset, cross-dataset, and zero-shot cross-dataset cross-lingual setups, reveal that fine-tuned deep neural Transformer-based BERT-like models outperform other machine learning methods in the task of automatic genre identification. The superior performance of these models is particularly evident in cross-dataset evaluation scenarios, where traditional non-neural machine learning approaches yield micro-F1 and macro-F1 scores of approximately 0.50 or below, indicating that they are not suitable for constructing a robust genre classifier. In contrast, the fine-tuned BERT-like model achieves remarkably high performance even on test sets using non-Latin scripts and on languages not closely related to those used during fine-tuning. These findings suggest that the final BERT-based genre classifier has strong potential for application across all languages included in the model's pretraining data.

Additionally, in light of recent advances in large language models (LLMs), we also evaluate the efficacy of the recently introduced zero-shot prompting approach using LLMs for the task of automatic genre identification (see Section 4.5 in Chapter 4). The findings indicate that both open-source and closed-source LLMs demonstrate robust zero-shot performance across all languages except Maltese, achieving macro-F1 scores as high as 0.87. The evaluation also shows that the test datasets, integrated into a newly developed benchmark, are not only useful for determining the state-of-the-art technology for automatic genre identification but also for evaluating the capabilities of large language models. The results on this benchmark highlight rapid advances in LLM development and significant differences in LLM performance across different languages.

Finally, leveraging the high performance of large language models in a zero-shot prompting setup, and the computational efficiency of the fine-tuned BERT-like models, we introduce a novel methodology for developing text classification models without the need for manually-annotated training data in Chapter 5. The proposed methodology, termed the LLM Teacher-Student Framework, uses a large language model as a data annotator – the teacher model – to automatically annotate training datasets with accuracy comparable to human annotation. A BERT-like model – the student – is then fine-tuned on this automatically-annotated data to produce a classifier that is more computationally efficient. Our evaluation, conducted on three test datasets across more than ten languages, shows that the student model achieves strong performance, thereby demonstrating that the LLM Teacher-Student Framework is a viable method for developing reliable genre classifiers. Second, the results indicate that this approach effectively leverages the strengths of both LLMs and BERT-like models: LLMs provide high-quality annotations through zero-shot prompting, while BERT-like models benefit from domain adaptation through fine-tuning on specific instances, which exposes them to specific characteristics of the training texts. This domain-specific fine-tuning enables the student model to occasionally even outperform the teacher model. However, it is important to acknowledge that the student model

rarely matches the performance of the same base model fine-tuned on human-annotated data. This underscores the added value of human annotation, which, despite being more time-consuming, labor-intensive, and costly, generally yields higher-quality training data. We do not argue that human annotation will become obsolete. Particularly for evaluation purposes, human-annotated test sets remain indispensable for ensuring reliable results. Nonetheless, in scenarios where large-scale annotation of multilingual training data is required, human annotation may be impractical due to resource constraints. In such instances, LLM-based annotation can be a highly effective alternative.

In summary, this thesis addresses the task of automatic genre identification from multiple complementary perspectives, with a strong emphasis on advancing the field while maintaining comparability with prior research. We introduce novel genre schemata that facilitate reliable manual annotation and the development of robust classifiers while remaining mappable to previously used schemata. This design choice directly addresses a key limitation of earlier work, namely the lack of comparability across studies due to incompatible genre schemata, and enables the reuse and joint evaluation of previously developed datasets. We improve the manual annotation methodology for the task of automatic genre identification to further increase the reliability of annotation campaigns. Building on these foundations, we develop several new training and test datasets in multiple European languages, which support the creation of robust genre classifiers and the evaluation of diverse machine learning methods on the task of automatic genre identification across different evaluation setups. The test datasets are integrated into a multilingual benchmark, enabling long-term evaluation of various models for this task and providing insight into the performance of large language models across different languages. Using this benchmark, we show that a model trained on only a small set of languages can achieve strong performance on a wide range of diverse and even unrelated languages. We also provide one of the early systematic evaluations of large language models for text classification in a zero-shot prompting scenario, demonstrating that high-quality genre classification can be achieved using only label definitions. Finally, motivated by the limitations of manual annotation, we introduce a novel teacher–student pipeline for developing genre classifiers without manually-annotated data. To the best of our knowledge, this approach has not previously been applied to this task and is potentially transferable to other text classification problems, thereby extending the impact of this work beyond automatic genre identification. The subsequent section further elaborates on the scientific contributions of this thesis.

6.1 Summary of Scientific Contributions

In this thesis, we address the entire process of developing automatic genre classifiers, from the formulation of a genre schema, the creation of a manually-annotated dataset, to machine learning experiments, and the validation of classifier robustness. The research provides significant contributions to the academic community, particularly in the areas of genre classification, text classification, multilingual natural language processing (NLP), and the development of language resources for less-resourced languages.

We introduce novel genre schemata that allow annotation with genre of the entire composition of not only web documents, but textual documents in general. Furthermore, these schemata have been demonstrated to support reliable manual annotation with high inter-annotator agreement. As detailed in Chapter 3, the GINCO and X-GENRE schemata meet the following essential criteria: they offer sufficient coverage to enable classification of the majority of texts in a web collection under specific labels; they comprise a limited number of labels, as increased granularity can negatively impact both the reliability of

manual annotations and the performance of models; and they are accessible to a broader audience through the use of labels with intuitive names. Through the development of these schemata, we achieve Goal *G1* "Develop a high-coverage genre schema that encompasses the full spectrum of genre variation in web texts and ensures reliable manual annotation with high inter-annotator agreement." The novel schemata are presented in Sections 3.1.1 and 3.2.1, which are based on the papers "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al., 2022) in Section 3.1.5, and "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4.

We propose a novel methodology for manual genre annotation that ensures high manual annotation reliability, fulfilling Goal *G2*. The improvements to the annotation methodology include the establishment of objective annotation criteria and the involvement of expert annotators who undergo continuous training. Significant effort is devoted to the development of comprehensive annotation guidelines to ensure the annotators' shared understanding of genre labels. The guidelines for both proposed schemata are publicly accessible to support future manual annotation campaigns.[1] Our findings demonstrate that the proposed manual annotation approach achieves acceptable inter-annotator agreement in genre identification annotation campaigns, surpassing both the threshold for acceptable inter-annotator agreement and the scores reported in previous studies (Egbert et al., 2015; Sharoff, 2018). These findings are consistent with Hypothesis *H1*: "*Acceptable inter-annotator agreement in genre identification annotation campaigns can be accomplished through a manual annotation approach that is based on a well-designed genre schema, comprehensive annotation guidelines, and employment of expert annotators.*". The methodology for manual genre annotation is presented in Section 3.1.2 in Chapter 3, based on the paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al., 2022) in Section 3.1.5.

We present the first Slovenian dataset manually-annotated with genre information: the Genre Identification Corpus (GINCO). This dataset facilitates the training and evaluation of machine learning models for automatic genre identification on the Slovenian language. We then explore the integration of the GINCO dataset with other English manually-annotated datasets, resulting in the development of a new Slovenian-English X-GENRE dataset. The X-GENRE dataset brings together three datasets in two languages, collected using different methods, sources, and time periods. This diversity improves the generalization and robustness of models trained on it. Furthermore, the inclusion of instances in both English and Slovenian languages permits the analysis of multilingual performance of genre classifiers. In adherence to the FAIR principles (Findable, Accessible, Interoperable, and Reusable), we make the GINCO dataset (Kuzman et al., 2021) and the X-GENRE dataset (Kuzman & Ljubešić, 2024a) freely accessible by depositing them in the CLARIN.SI and Hugging Face repositories to ensure their long-term preservation and open access.[2] The development of these training datasets, which fulfills the first part of Goal *G3*, and the examination of their comparability with other existing datasets are detailed in Sections 3.1.3, 3.2, and 3.2.2, and in the following papers: "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al., 2022) in Section

[1]The GINCO annotation guidelines can be accessed at https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/ and the X-GENRE annotation guidelines are available at https://tajakuzman.github.io/X-GENRE-Annotation-Guidelines/.

[2]The X-GENRE dataset is accessible on the CLARIN.SI repository at http://hdl.handle.net/11356/1960 and the Hugging Face repository at https://doi.org/10.57967/hf/6582, and the GINCO dataset is available on the CLARIN.SI repository at http://hdl.handle.net/11356/1467 and the Hugging Face repository at https://huggingface.co/datasets/cjvt/ginco.

3.1.5, "*Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments*" (Kuzman, Ljubešić, & Pollak, 2022) in Section 3.2.3, and "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4.

We develop test datasets for a variety of European languages, thereby fulfilling Goal *G3*: "Develop a genre dataset for Slovenian, which will be the first of its kind for this language, to enable training and evaluation of machine learning models in automatic genre identification on Slovenian language. Additionally, we will construct training datasets that include English, and separate test datasets for various European languages to assess the performance of genre classifiers in cross-dataset and cross-lingual evaluation scenarios." The newly developed EN-GINCO and X-GINCO test datasets encompass eleven languages from three major language families – Indo-European, Turkic, and Afro-Asiatic – and include different writing systems, thus enabling a comprehensive evaluation of genre classifiers across both languages and scripts. Building upon these datasets, we establish the AGILE Benchmark (Automatic Genre Identification Benchmark), freely available on GitHub,[3] thereby fulfilling Goal *G4*. The benchmark addresses the lack of reference evaluation datasets and enables comparative analyses of machine learning techniques employed in the task of automatic genre identification. Additionally, our experiments with large language models (LLMs) demonstrate that the AGILE benchmark serves as a valuable resource not only for assessing the state of the art in automatic genre identification but also for evaluating and comparing large language models across languages. We integrate the EN-GINCO and X-GINCO test datasets into a broader benchmark for LLM evaluation that encompasses a range of text classification and commonsense reasoning tasks for South Slavic languages (Kuzman Pungeršek, Rupnik, Porupski, et al., 2026). The results of the evaluations on this benchmark are also published in the form of an interactive dashboard, the "CLASSLA LLM Evaluation Dashboard for South Slavic Languages",[4] which enables interactive exploration and analysis of LLM evaluation results across languages and tasks. The development of the test datasets is detailed in Section 4.1.1 in Chapter 4, and in the papers "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4 and "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al., In submission) in Section 4.6.

We conduct experiments with various machine learning techniques to identify the most reliable methodology for genre classification. Our findings indicate that the BERT-like XLM-RoBERTa model, fine-tuned on the English-Slovenian training dataset, achieves state-of-the-art performance on this task and is capable of generalizing across different datasets and European languages. The results of these experiments provide empirical support for Hypothesis *H2* "*Fine-tuned deep neural Transformer-based language models outperform other machine learning methods in the task of automatic genre identification in the Slovenian language.*" and Hypothesis *H3* "*A reliable zero-shot cross-lingual performance in automatic genre identification across numerous European languages is possible by leveraging the cross-lingual capabilities of massively multilingual pre-trained Transformer models.*" By using the best-performing machine learning method and the most suitable training dataset, we develop a robust multilingual genre classifier, termed the X-GENRE classifier, thereby achieving Goal *G5*. The X-GENRE classifier (Kuzman & Ljubešić, 2024c) has been made available on the CLARIN.SI repository and the Hugging Face repository.[5] The

[3]https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

[4]https://www.clarin.si/classla-llm-dashboard/

[5]The X-GENRE classifier is freely accessible on the CLARIN.SI repository at http://hdl.handle.net/

machine learning experiments are detailed in Sections 4.2, 4.3, 4.4, and 4.5, and in the papers "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al., 2023) in Section 3.2.4, "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al., In submission) in Section 4.6, and "*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?*" (Kuzman Pungeršek, Rupnik, Porupski, et al., 2026) in Section 4.7.

Leveraging recent advancements in large language model capabilities, we propose an innovative methodology for developing text classifiers without the need for manually-annotated data, termed the LLM Teacher-Student Framework. Our findings indicate that the LLM annotator surpasses the threshold for minimally acceptable inter-annotator agreement. Furthermore, the agreement between the LLM annotator and human annotators is significantly higher than the human inter-annotator agreement reported in related genre annotation studies (Egbert et al., 2015; Sharoff, 2018). These results suggest that the LLM is a viable data annotator. We show that replacing human annotators with a large language model for training data annotation is an effective strategy for developing robust and efficient genre classifiers. The methodology is applicable not only to genre identification, but also to other text classification tasks, such as topic classification. The novel methodology is detailed in Chapter 5 and published in the paper titled *"LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification"* (Kuzman & Ljubešić, 2025), included in Section 5.3.

Although applying the developed genre classifier to real-world annotation scenarios is beyond the scope of this thesis, our connected work confirms the practical utility of the developed classifier. Inter alia, we applied the X-GENRE classifier to large web corpora in ten less-resourced European languages (see Kuzman Pungeršek et al. (In submission) in Section 4.6), resulting in a multilingual genre-annotated corpus collection of nearly 53 million texts. These corpora, released as the MaCoCu-Genre dataset (Kuzman & Ljubešić, 2024b), are openly available on the CLARIN.SI repository.[6] The X-GENRE classifier has also been used to annotate parallel MaCoCu corpora aligned with English texts (Kuzman, Rupnik, et al., 2023), and two releases of the CLASSLA-web corpora (Kuzman Pungeršek, Rupnik, Suchomel, et al., 2026; Ljubešić & Kuzman, 2024) that encompass the entire South Slavic language family and total 57 million unique texts.[7] Given their scale and recency, these massive multilingual corpora are key resources for developing language technologies and conducting linguistic analyses for the less-resourced languages that they cover. They have already been used for pretraining BERT-like and decoder-only large language models (Ljubešić, Suchomel, et al., 2024; Vreš et al., 2024), for building datasets for downstream NLP tasks (Kuzman & Ljubešić, 2025; Kuzman et al., 2024), and have been shown to be heavily consulted by linguists (Erjavec et al., 2024; Ljubešić, Kuzman, et al., 2024). To encourage broader use among linguists, language teachers, and digital humanists, we introduced the genre-annotated CLASSLA-web corpora in the CLASSLA-Express workshop series (Ljubešić, Kuzman, et al., 2024), held in eleven cities across six countries in 2024 and 2025. The workshops showed participants how to use the CLASSLA-web corpora in language research, which includes linguistic analyses that leverage genre information.

Genre annotations have also proven useful for exploring similarities and differences between the corpora. For example, a genre-based analysis of CLASSLA-web 1.0 (Ljubešić & Kuzman, 2024) showed consistent genre distributions across the seven South Slavic

11356/1961 and the Hugging Face repository at https://doi.org/10.57967/hf/0927.

[6]The MaCoCu-Genre dataset is freely available at http://hdl.handle.net/11356/1969.

[7]The CLASSLA-web corpora are available on the CLARIN.SI repository and can be queried through the CLARIN.SI concordancer. The links to each of the corpora are available at https://clarinsi.github.io/classla-web/.

corpora, with variations aligning with economic differences. We found that web corpora from economically less developed countries are dominated by news content, while those from more developed regions include more promotional and opinionated texts. Similarly, in Kuzman, Rupnik, et al. (2023), analysis of seven parallel MaCoCu corpora, covering Bulgarian, Croatian, Icelandic, Macedonian, Maltese, Slovenian, and Turkish, revealed important differences between them. For instance, the Macedonian, Maltese, and Icelandic corpora were shown to comprise more news articles than other corpora, whereas the Slovenian, Croatian, Bulgarian, and Turkish corpora contain more promotional texts. These applications highlight the value of automatic genre annotation for understanding corpus composition and cross-linguistic variation.

Additionally, the genre-annotated text collections support the development of genre-specific datasets for corpus linguistics and various natural language processing tasks. The genre-enriched datasets have already been used in multiple studies that benefit from genre-based data selection. In a study focused on the development of a news topic classification model, news articles were extracted from the MaCoCu-Genre dataset by filtering for the *News* genre label (see Kuzman and Ljubešić (2025) in Section 5.3). Another of our studies used the Slovenian MaCoCu-Genre corpora and a selection of genre labels to develop a question-answering test dataset for evaluating retrieval-augmented generation systems in Slovenian (Kuzman et al., 2024).

6.2 Further Work

There are several promising directions for extending this work and addressing its current limitations. First, we could further improve the performance of both fine-tuned BERT-like models and large language models (LLMs). This could involve expanding the X-GENRE training dataset by (a) incorporating additional instances from the CORE (Egbert et al., 2015) and FTD (Sharoff, 2018) datasets – they have been only partially included in the X-GENRE dataset to ensure a balanced distribution of instances across the three datasets, and (b) including additional languages by mapping the X-GENRE schema to existing annotated corpora using the FTD or CORE schema, such as the Russian FTD dataset (Sharoff, 2018) and Finnish, Swedish, and French datasets based on the CORE schema (Laippala et al., 2020; Skantsi & Laippala, 2023). The evaluation could also be extended to include languages that are more typologically and geographically diverse. While our current work covers over ten languages, including several that are less-resourced, they are all European. Future research could address this by incorporating languages such as Mandarin Chinese, Hindi, and Arabic, thereby broadening the scope of multilingual genre classification.

Further work could also explore the potential of newer multilingual Transformer-based encoder models. While XLM-RoBERTa (Conneau et al., 2020) remains a strong baseline for automatic genre identification, recently introduced models like mmBERT, pretrained on over 1,800 languages, have been reported to improve both performance and speed, particularly for low-resource languages (Marone et al., 2025). It remains to be seen whether such models would offer improvements in genre classification over the XLM-RoBERTa model. The performance of LLMs could also be enhanced through more sophisticated prompting strategies. Techniques such as chain-of-thought prompting (Wei et al., 2022) could be explored, as well as experimenting with different genre schemata used in the prompts, namely, the CORE (Egbert et al., 2015), FTD (Sharoff, 2018), or GINCO schemata. Another possible direction is to investigate the effectiveness of fine-tuning LLMs using parameter-efficient methods like low-rank adaptation (LoRA; Hu et al. (2021)) and QLoRA (Dettmers et al., 2023), which allow large models to be adapted with significantly lower memory require-

ments.

Maltese was shown to be a particularly challenging case in our evaluations for both BERT-like models and LLMs (see Sections 4.4 and 4.5). One possible explanation is its limited representation in the pretraining data of the evaluated Transformer-based models. However, this interpretation requires further empirical validation. Previous studies on other tasks have shown that additional pretraining on monolingual data can substantially improve performance for under-represented languages such as Maltese (Chau et al., 2020; Ebrahimi & Kann, 2021; Muller et al., 2021). Future work could examine whether performance on languages that are not included in the model's pretraining data could be improved by additional pretraining, or by fine-tuning on genre-annotated data prepared specifically in these languages. If the performance of large language models continues to improve on under-represented languages, this could also open up the possibility of using LLM-based annotation to create genre-labeled training data for Maltese and other languages on which the current state-of-the-art technologies underperform.

Building robust genre classifiers for single-label classification is a crucial first step toward reliable automatic genre identification. However, due to the limited control over content creation on the web, previous work has pointed to the need for solutions for dealing with complex cases, such as hybrid texts, i.e., texts that have features of multiple genres, texts without any discernible purpose or features, and documents that consist of multiple texts, e.g., a letter, posted as a part of a blog (Kwaśnik & Crowston, 2005; Repo et al., 2021; Sharoff, 2021). Although the GINCO dataset was constructed with such cases in mind, as it includes some multi-label annotations and instances of multi-text documents, these challenges were not explored in depth in our current work. Future work could focus on multi-label classification, span-level genre annotation within documents, and methods for handling noisy documents.

Explainability is another area that could be explored in more detail. While not the focus of the thesis, we have explored the role of lexical and syntactic features in genre classification using fastText (Kuzman & Ljubešić, 2022), and examined the features that impact the cross-lingual model performance (Kuzman Pungeršek et al., In submission) in our related works. These studies suggest that syntactic features may play a significant role in genre differentiation, but the findings are based on small-scale experiments. Scaling up these analyses to more languages, genre categories, and instances would yield more robust conclusions. It would also be insightful to explore whether the characteristics of genre labels are consistent across cultures and languages.

Finally, the high computational costs associated with Transformer-based models remain a key limitation. Both BERT-like encoders and decoder-only LLMs require significant hardware resources for training and inference, which make them less accessible to individuals, researchers, and organizations with resource constraints. While this issue falls outside the scope of the current work, future research could focus on developing more efficient architectures that reduce computational requirements while maintaining performance. Addressing this challenge could make genre classification tools even more widely usable.

6.3 Implementation and Availability

The source code for the experiments presented in this dissertation is publicly available under an MIT license to ensure the reproducibility of the conducted research. Specifically, the code can be found in the following GitHub repositories:

- The code for experiments evaluating the suitability of the GINCO schema and dataset for the development of genre classifiers, presented in Section 3.1.4 and described in

the paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al. (2022); Section 3.1.5):

https://github.com/TajaKuzman/GINCO-web-genre-experiments-paper

- The code for experiments exploring the integration of the Slovenian GINCO dataset with pre-existing English manually-annotated genre datasets, presented in Section 3.2 and described in the paper "*Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments*" (Kuzman, Ljubešić, and Pollak (2022); Section 3.2.3):

  https://github.com/TajaKuzman/Cross-Lingual-and-Cross-Dataset-Experiments-with-Genre-Datasets

- The code for the development of the joint X-GENRE schema, and for machine learning experiments using different genre schemata, genre datasets and machine learning methods, presented in Sections 3.2.1, 3.2.2, 4.2 and 4.3, and described in the paper "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al. (2023); Section 3.2.4):

  https://github.com/TajaKuzman/X-GENRE-Classifier-Development

- The code for cross-lingual zero-shot experiments, presented in Section 4.4 and described in the paper "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al. (In submission); Section 4.6):

  https://github.com/TajaKuzman/Zero-Shot-Cross-Lingual-Genre-Experiments

- The code for experiments evaluating the performance of large language models on the automatic genre identification benchmarks, presented in Section 4.5 and described in the paper "*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?*" (Kuzman Pungeršek, Rupnik, Porupski, et al. (2026); Section 4.7):

  https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

- The code for experiments based on the LLM Teacher-Student framework for the development of a genre classifier without manually-annotated data, presented in Chapter 5 and described in the paper *"LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification"* (Kuzman and Ljubešić (2025); Section 5.3):
  - https://github.com/TajaKuzman/IPTC-Media-Topic-Classification (experiments with news topic classification)
  - https://github.com/TajaKuzman/LLM-Teacher-Student-Framework-on-Genre (experiments with genre identification)

Adhering to the FAIR principles, we have made available manually-annotated genre datasets, automatically-annotated web corpora, and trained genre classifiers, namely:

- The Slovenian manually-annotated genre dataset GINCO (Kuzman et al., 2021), presented in Section 3.1.3 and described in the paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik,

et al. (2022); Section 3.1.5), is freely available under the CC BY-SA 4.0 license on the CLARIN.SI repository at http://hdl.handle.net/11356/1467 and on the Hugging Face repository at https://huggingface.co/datasets/cjvt/ginco.

- The English-Slovenian manually-annotated genre dataset X-GENRE (Kuzman & Ljubešić, 2024a), presented in Section 3.2.2, and described in the paper "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al. (2023); Section 3.2.4), is freely available under the CC BY-SA 4.0 license on the CLARIN.SI repository at http://hdl.handle.net/11356/1960, and on the Hugging Face repository at https://doi.org/10.57967/hf/6582.

- The multilingual genre classifier X-GENRE (Kuzman & Ljubešić, 2024c) – the best performing XLM-RoBERTa model, fine-tuned on the X-GENRE dataset – presented in Section 4.8, and described in the paper "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models*" (Kuzman, Mozetič, et al. (2023), Section 3.2.4), is freely available under the CC BY-SA 4.0 license on the CLARIN.SI repository at http://hdl.handle.net/11356/1961, and on the Hugging Face repository at https://doi.org/10.57967/hf/0927.

- The MaCoCu-Genre corpora (Kuzman & Ljubešić, 2024b) that have been automatically annotated with genres with the X-GENRE classifier, mentioned in Sections 4.1.1 and 6.1 and described in the paper "*Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages*" (Kuzman Pungeršek et al. (In submission); Section 4.6) are freely available under the CC0 license on the CLARIN.SI repository at http://hdl.handle.net/11356/1969.

- The CLASSLA-web corpora that have been automatically annotated with genres with the X-GENRE classifier, mentioned in Section 6.1, are freely available under the CC0 license on the CLARIN.SI repository at https://www.clarin.si/repository/xmlui/ and can be queried through the CLARIN.SI concordancer at https://www.clarin.si/ske/. The links to each of the corpora are available at https://clarinsi.github.io/classla-web/.

The EN-GINCO and X-GINCO test datasets have not been published online to prevent their inclusion in the training data of large language models. This measure facilitates reliable long-term evaluation on these benchmarks. Instead, the datasets are available upon request from the authors and have been integrated into the AGILE Benchmark (Automatic Genre Identification Benchmark), which is publicly available on GitHub: https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark.

Additionally, we have published the genre annotation guidelines to facilitate future genre annotation campaigns. They can be accessed at the following websites:

- The genre annotation guidelines for the GINCO schema, presented in Sections 3.1.1 and 3.1.2, and described in the paper "*The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild*" (Kuzman, Rupnik, et al. (2022); Section 3.1.5), are published at https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

- The genre annotation guidelines for the X-GENRE schema, presented in Section 3.2.1 and described in the paper "*Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large*

Language Models" (Kuzman, Mozetič, et al. (2023); Section 3.2.4), are published at https://tajakuzman.github.io/X-GENRE-Annotation-Guidelines/.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. https://arxiv.org/abs/2303.08774

Agrawal, S., Sanagavarapu, L. M., & Reddy, Y. R. (2019). FACT-Fine grained Assessment of web page CredibiliTy. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 1088–1097.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. *First International workshop on innovative information systems*, 85–92.

Asheghi, N. R., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, *50*(3), 603–641.

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., et al. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. *23rd Annual Conference of the European Association for Machine Translation*, 301–302.

Bañón, M., Esplà-Gomis, M., Forcada, M. L., Garcıéa-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Pla Sempere, L., Ramıérez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., & Zaragoza, J. (2022). Slovene web corpus MaCoCu-sl 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1517

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Berninger, V. F., Kim, Y., & Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. *BCS-IRSG Workshop on Corpus Profiling*, 1–10.

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press.

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Boese, E. S. (2005). *Stereotyping the web: Genre classification of web documents* (Doctoral dissertation). Citeseer.

Chandler, D. (1997). An introduction to genre theory.

Chau, E. C., Lin, L. H., & Smith, N. A. (2020). Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1324–1334.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. https://arxiv.org/abs/2507.06261

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Crowston, K., Kwaśnik, B., & Rubleske, J. (2010). Problems in the use-centered development of a taxonomy of web genres. *Genres on the Web* (pp. 69–84). Springer.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, *36*(1), 1–28.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, *36*, 10088–10115.

de Vries, W., Wieling, M., & Nissim, M. (2022). Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. *The 60th Annual Meeting of the Association for Computational Linguistics*, 7676–7685.

Dewe, J., Karlgren, J., & Bretan, I. (1998). Assembling a balanced corpus from the internet. *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, 100–108.

Ebrahimi, A., & Kann, K. (2021). How to adapt your pretrained multilingual model to 1600 languages. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4555–4567.

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, *66*(9), 1817–1831.

Erjavec, T., & Ljubešić, N. (2014). The slWaC 2.0 corpus of the Slovene web. *T. Erjavec, J. Žganec Gros (ur.) Jezikovne tehnologije: zbornik*, *17*, 50–55.

Erjavec, T., Ljubešić, N., Meden, K., Kuzman, T., Laskowski, C. A., Javoršek, J. J., Krek, S., Jemec Tomazin, M., & Lenardič, J. (2024). CLARIN.SI, the Slovenian node of CLARIN: ten years on. *Proceedings of the CLARIN 2024 Conference*, 76–80.

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, *57*(11), 1506–1518.

Forsyth, R. S., & Sharoff, S. (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*, *29*(1), 6–22.

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*. https://arxiv.org/pdf/2503.19786

Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. *Fifth Web as Corpus Workshop*, 27–35.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*. https://arxiv.org/pdf/2501.12948?

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210*.

Henriksson, E., Myntti, A., Hellström, S., Eskelinen, A., Erten-Johansson, S., & Laippala, V. (2024). Automatic register identification for the open web using multilingual deep learning. *arXiv preprint arXiv:2406.19892*. https://arxiv.org/abs/2406.19892

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2021). Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *Companion proceedings of the ACM web conference 2023*, 294–297.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The TenTen corpus family. *7th international corpus linguistics conference CL*, 125–127.

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Kilgarriff, A. (2012). Getting to know your corpus. *International conference on text, speech and dialogue*, 3–15.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kuzman, T., Brglez, M., Rupnik, P., & Ljubešić, N. (2021). Slovene web genre identification corpus GINCO 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1467

Kuzman, T., & Ljubešić, N. (2022). Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In D. Mladenić & M. Grobelnik (Eds.), *Data Mining and Data Warehouses - SiKDD, 10 October 2022, Ljubljana, Slovenia*. Jožef Stefan Institute.

Kuzman, T., & Ljubešić, N. (2023). Automatic genre identification: A survey. *Language Resources and Evaluation*, 1–34. https://doi.org/10.1007/s10579-023-09695-8

Kuzman, T., & Ljubešić, N. (2024a). English-Slovenian text genre dataset X-GENRE [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1960

Kuzman, T., & Ljubešić, N. (2024b). Genre-enriched web corpora MaCoCu-Genre [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1969

Kuzman, T., & Ljubešić, N. (2024c). Multilingual text genre classification model X-GENRE [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1961

Kuzman, T., & Ljubešić, N. (2025). LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, *13*, 35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814

Kuzman, T., Ljubešić, N., & Pollak, S. (2022). Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the conference on language technologies and digital humanities* (pp. 100–107). Institute of Contemporary History.

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in

the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149–1175. https://doi.org/10.3390/make5030059

Kuzman, T., Pavleska, T., Rupnik, U., & Cigoj, P. (2024). PandaChat-RAG: Towards the benchmark for Slovenian RAG Applications. *Proceedings of the Slovenian Conference on Artificial Intelligence 2024, Vol. A, Ljubljana, Slovenia*, 7–10.

Kuzman, T., Rupnik, P., & Ljubešić, N. (2022). The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. *Language Resources and Evaluation Conference*, 1584–1594.

Kuzman, T., Rupnik, P., & Ljubešić, N. (2023). Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 91–103.

Kuzman Pungeršek, T., Mozetič, I., & Ljubešić, N. (In submission). Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages. *Natural Language Processing*.

Kuzman Pungeršek, T., Rupnik, P., Porupski, I., Dinić, V., & Ljubešić, N. (2026). State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting? *Proceedings of the LLMs4SSH 2026 workshop, in conjunction with LREC 2026*.

Kuzman Pungeršek, T., Rupnik, P., Suchomel, V., & Ljubešić, N. (2026). The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora. *International Conference on Language Resources and Evaluation (LREC 2026)*.

Kwaśnik, B. H., & Crowston, K. (2005). Introduction to the special issue: Genres of digital documents. *Information technology & people*.

Laippala, V., Egbert, J., Biber, D., & Kyröläinen, A.-J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*, 1–32.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. *22nd Nordic Conference on Computational Linguistics*, 292–297.

Laippala, V., Luotolahti, J., Kyröläinen, A.-J., Salakoski, T., & Ginter, F. (2017). Creating register sub-corpora for the Finnish Internet Parsebank. *21st Nordic Conference on Computational Linguistics*, 152–161.

Laippala, V., Rönnqvist, S., Hellström, S., Luotolahti, J., Repo, L., Salmela, A., Skantsi, V., & Pyysalo, S. (2020). From web crawl to clean register-annotated corpora. *12th Web as Corpus Workshop*, 14–22.

Laippala, V., Rönnqvist, S., Oinonen, M., Kyröläinen, A.-J., Salmela, A., Biber, D., Egbert, J., & Pyysalo, S. (2022). Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 1–35.

Lee, D. (2002). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Teaching and Learning by Doing Corpus Analysis* (pp. 245–292). Brill Rodopi.

Lee, Y.-B., & Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 145–150.

Lepekhin, M., & Sharoff, S. (2022). Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for the Genre Classification Task. *Language Resources and Evaluation Conference*, 5974–5982.

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information processing & management*, *41*(5), 1263–1276.

Ljubešić, N., Galant, N., Benčina, S., Čibej, J., Milosavljević, S., Rupnik, P., & Kuzman, T. (2024). DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects. *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 89–98.

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282.

Ljubešić, N., Kuzman, T., Filipović Petrović, I., Parizoska, J., & Osenova, P. (2024). CLASSLA-Express: a Train of CLARIN. SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings 2024*, 31–35.

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, 189–203.

Maeda, A., & Hayashi, Y. (2009). Automatic genre classification of Web documents using discriminant analysis for feature selection. *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, 405–410.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., & Van Durme, B. (2025). mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. *arXiv preprint arXiv:2509.06888*.

Meta. (2024). Llama 3.3 Model Card [Accessed: June 26, 2025]. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

Mistral AI. (2025). Medium is the new large. [Accessed: October 10, 2025]. https://mistral.ai/news/mistral-medium-3

Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 448–462.

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021). Genre as Weak Supervision for Cross-lingual Dependency Parsing. *2021 Conference on Empirical Methods in Natural Language Processing*, 4786–4802.

OpenAI. (2023). ChatGPT General FAQ [Accessed: March 3, 2023]. https://help.openai.com/en/articles/6783457-chatgpt-general-faq

OpenAI. (2024). Hello GPT-4o [Accessed September 11, 2024]. https://openai.com/index/hello-gpt-4o/

OpenAI. (2025). Introducing GPT-5 [Accessed: October 10, 2025]. https://openai.com/index/introducing-gpt-5/

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, 541–574.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., & Launay, J. (2023). The RefinedWeb dataset for Falcon LLM:

Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, *36*, 79155–79172.

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics*, *37*(2), 385–393.

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., & Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. *Proc. 8th Web as Corpus Workshop (WAC-8)*, 24–29.

Pritsos, D., & Stamatatos, E. (2018). Open set evaluation of web genre identification. *Language Resources and Evaluation*, *52*(4), 949–968.

Priyatam, P. N., Iyengar, S., Perumal, K., & Varma, V. (2013). Don't Use a Lot When Little Will Do: Genre Identification Using URLs. *Res. Comput. Sci.*, *70*, 233–243.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., & Vidulin, V. (2008). Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. *LREC*.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo, S., & Laippala, V. (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021*, 183–191.

Rezapour Asheghi, N. (2015). *Human annotation and automatic detection of web genres* (Doctoral dissertation). University of Leeds.

Rönnqvist, S., Skantsi, V., Oinonen, M., & Laippala, V. (2021). Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification. *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 157–165.

Rosso, M. A. (2008). User-based identification of Web genres. *Journal of the American Society for Information Science and Technology*, *59*(7), 1053–1072.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., & Liu, X. (2001). Genre based navigation on the web. *34th annual Hawaii international conference on system sciences*, 10–pp.

Santini, M. (2007). *Automatic identification of genre in web pages* (Doctoral dissertation). University of Brighton.

Santini, M. (2010). Cross-testing a genre classification model for the web. *Genres on the Web* (pp. 87–128). Springer.

Santini, S. M. (2006). Common criteria for genre classification: Annotation and granularity. *Workshop on Text-based Information Retrieval (TIR-06), In Conjunction with ECAI 2006, Riva del Garda, 2006*.

Sharoff, S. (2010). In the Garden and in the Jungle. *Genres on the Web* (pp. 149–166). Springer.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, *13*(1), 65–95.

Sharoff, S. (2021). Genre annotation for the web: Text-external and text-internal perspectives. *Register studies*.

Sharoff, S., Wu, Z., & Markert, K. (2010). The Web Library of Babel: evaluating genre collections. *LREC*.

Shavrina, T. (2019). Genre classification problem: In pursuit of systematics on a big web-corpus. *Proceedings of Third Workshop" Compu*, *4*, 70–83.

Skantsi, V., & Laippala, V. (2023). Analyzing the unrestricted web: The Finnish corpus of online registers. *Nordic Journal of Linguistics*, 1–31.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics*, *26*(4), 471–495.

Stein, B., Eissen, S. M. z., & Lipka, N. (2010). Web genre analysis: Use cases, retrieval models, and implementation issues. *Genres on the Web* (pp. 167–189). Springer.

Stewart, J. G., & Callan, J. (2009). *Genre oriented summarization* (Doctoral dissertation). Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Stubbe, A., & Ringlstetter, C. (2007). Recognizing genres. *Proc. Towards a Reference Corpus of Web Genres.*

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture.* Blackwell Oxford.

Suchomel, V. (2020). Genre Annotation of Web Corpora: Scheme and Issues. *Future Technologies Conference*, 738–754.

Ulčar, M., & Robnik-Šikonja, M. (2021a). SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD.*

Ulčar, M., & Robnik-Šikonja, M. (2021b). Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1397

Van der Wees, M., Bisazza, A., & Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. *Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Vidulin, V., Luštrek, M., & Gams, M. (2007). Using genres to improve search engines. *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, 45–51.

Vreš, D., Božič, M., Potočnik, A., Martinčič, T., & Robnik Šikonja, M. (2024). Generative model for less-resourced language with 1 billion parameters. *Jezikovne tehnologije in digitalna humanistika*, 485–511.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems.*

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Williams, M., Kevin Crowston. (2000). Reproduced and emergent genres of communication on the World Wide Web. *The information society*, *16*(3), 201–215.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388.* https://arxiv.org/abs/2505.09388

Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. *Thirty-fourth International Conference on Machine Learning (ICML 2017).*

Zhang, B., Ding, D., & Jing, L. (2022). How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548.*

Zu Eissen, S. M., & Stein, B. (2004). Genre classification of web pages. *Annual Conference on Artificial Intelligence*, 256–269.

Bibliography

Publications Related to the Thesis

Journal Articles

Kuzman, T., & Ljubešić, N. (2023). Automatic genre identification: A survey. *Language Resources and Evaluation*, 1–34. https://doi.org/10.1007/s10579-023-09695-8

Kuzman, T., & Ljubešić, N. (2025). LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, *13*, 35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149–1175. https://doi.org/10.3390/make5030059

Kuzman Pungeršek, T., Mozetič, I., & Ljubešić, N. (In submission). Zero-Shot Cross-Lingual Genre Classification on Diverse European Languages. *Natural Language Processing*.

Conference Papers

Kuzman, T. (2023). Troubling times for PhD research on text categorization?: ChatGPT for automatic genre identification. *ESSLLI 2023: 34th European summer school in logic, language and information*, 1–12.

Kuzman, T., & Ljubešić, N. (2022). Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In D. Mladenić & M. Grobelnik (Eds.), *Data Mining and Data Warehouses - SiKDD, 10 October 2022, Ljubljana, Slovenia*. Jožef Stefan Institute.

Kuzman, T., Ljubešić, N., & Pollak, S. (2022). Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the conference on language technologies and digital humanities* (pp. 100–107). Institute of Contemporary History.

Kuzman, T., Rupnik, P., & Ljubešić, N. (2022). The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. *Language Resources and Evaluation Conference*, 1584–1594.

Kuzman, T., Rupnik, P., & Ljubešić, N. (2023). Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 91–103.

Kuzman Pungeršek, T., Rupnik, P., Porupski, I., Dinić, V., & Ljubešić, N. (2026). State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting? *Proceedings of the LLMs4SSH 2026 workshop, in conjunction with LREC 2026*.

Kuzman Pungeršek, T., Rupnik, P., Porupski, I., & Ljubešić, N. (2025). Towards benchmarking text classification for South Slavic languages in the era of large language models. *Proceedings of the 1st ReLDI Conference on Language Science and Technology*, 59–62.

Kuzman Pungeršek, T., Rupnik, P., Suchomel, V., & Ljubešić, N. (2026). The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora. *International Conference on Language Resources and Evaluation (LREC 2026)*.

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282.

Other Publications

Journal Articles

Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., et al. (2024). ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 1–32.

Gantar, P., Arhar Holdt, Š., Čibej, J., & Kuzman, T. (2019). Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. *Contributions to Contemporary History*, *59*(1), 99–119.

Kosem, I., Čibej, J., Dobrovoljc, K., Kuzman, T., & Ljubešić, N. (2023). Spremljevalni korpus Trendi in avtomatska kategorizacija. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 161–188.

Mochtak, M., Rupnik, P., Kuzman, T., & Ljubešić, N. (2025). Parlasent: Mapping sentiment in political discourse with large language models. *Political Research Exchange*, *7*(1), 2508377.

Conference Papers

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., et al. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. *23rd Annual Conference of the European Association for Machine Translation*, 301–302.

de Jong, A., Kuzman, T., Larooij, M., & Marx, M. (2024). ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments. *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*, 110–115.

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022). Raziskovalna infrastruktura CLARIN.SI. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the conference on language technologies and digital humanities*. Institute of Contemporary History.

Erjavec, T., Ljubešić, N., Meden, K., Kuzman, T., Laskowski, C. A., Javoršek, J. J., Krek, S., Jemec Tomazin, M., & Lenardič, J. (2024). CLARIN.SI, the Slovenian node of CLARIN: ten years on. *Proceedings of the CLARIN 2024 Conference*, 76–80.

Esplà-Gomis, M., Forcada, M. L., Kuzman, T., Ljubešić, N., Van Noord, R., Ramıérez-Sánchez, G., Tiedemann, J., & Toral, A. (2023). Proceedings of the 1st Workshop on Open Community-Driven Machine Translation. *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, 1–2.

Gantar, P., Holdt, Š. A., Čibej, J., Kuzman, T., & Kavčič, T. (2018). Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. *Proceedings of the conference on Language Technologies & Digital Humanities*, 85–92.

Gantar, P., Krek, S., & Kuzman, T. (2017). Verbal multiword expressions in Slovene. *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, November 13-14, 2017, Proceedings 2*, 247–259.

Kuzman, T., & Fišer, D. (2017). Corpus-Based Analysis of Demonyms in Slovene Twitter. *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, 30.

Kuzman, T., Pavleska, T., Rupnik, U., & Cigoj, P. (2024). PandaChat-RAG: Towards the benchmark for Slovenian RAG Applications. *Proceedings of the Slovenian Conference on Artificial Intelligence 2024, Vol. A, Ljubljana, Slovenia*, 7–10.

Kuzman, T., Vintar, Š., & Arcan, M. (2019). Neural machine translation of literary texts from English to Slovene. *Proceedings of the qualities of literary machine translation*, 1–9.

Ljubešić, N., Galant, N., Benčina, S., Čibej, J., Milosavljević, S., Rupnik, P., & Kuzman, T. (2024). DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects. *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 89–98.

Ljubešić, N., Kuzman, T., Filipović Petrović, I., Parizoska, J., & Osenova, P. (2024). CLASSLA-Express: a Train of CLARIN. SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings 2024*, 31–35.

Ljubešić, N., Kuzman, T., Rupnik, P., Vulić, I., Schmidt, F., & Glavaš, G. (2024). JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far. *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 209–219.

Ljubešić, N., Porupski, I., Rupnik, P., & Kuzman, T. (2025). Identifying Filled Pauses in Speech Across South and West Slavic Languages. *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, 1–8.

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, 189–203.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešic, N., Çöltekin, Ç., Kopp, M., Meden, K., & Kuzman, T. (2023). The ParlaMint Project: Ever-growing Family of Comparable and Interoperable Parliamentary Corpora. *CLARIN Annual Conference Proceedings*, 62.

Rupnik, P., Kuzman, T., & Ljubešić, N. (2023). BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 113–120.

van Noord, R., Kuzman, T., Rupnik, P., Ljubešić, N., Esplà-Gomis, M., Ramıérez-Sánchez, G., & Toral, A. (2024). Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across 11 Languages. *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 5221–5234.

Biography

Taja Kuzman Pungeršek is a data scientist at International Flavors & Fragrances Inc. and an assistant researcher at the Department of Knowledge Technologies at the Jožef Stefan Institute in Ljubljana, Slovenia. She obtained a Master's degree in Translation between English, Slovenian, and French from the Faculty of Arts at the University of Ljubljana in 2019. Her Master's thesis focused on a topic from the computational linguistics field, namely, an evaluation of neural machine translation systems for literary texts. Since 2021, she has been undertaking doctoral studies in Information and Communication Technologies at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia.

Her research interests involve the collection and curation of language resources, as well as the development of technologies for automatic annotation using deep learning. The latter includes a broad range of token and text classification tasks, such as automatic genre identification, topic detection, hate speech classification, and named entity recognition. She was engaged in numerous EU-funded and Slovenian projects. Notable projects include the development of massive web text collections for various less-resourced European languages (MaCoCu project); the collection and curation of comparable parliamentary text collections from 29 European countries and regions, enriched with sentiment and topic information (ParlaMint II and ParlaCAP projects); the development of speech resources and technologies for Slovenian (Mezzanine project); the application of natural language processing techniques to news media monitoring (EMMA project); and the evaluation of large language models in South Slavic languages (LLM4DH and LLMs4EU projects).

Additionally, Ms. Kuzman Pungeršek served as a co-leader of the CLASSLA knowledge centre for South Slavic languages, providing expertise in language resources and technologies for these languages. Within CLASSLA, she was involved in benchmarking state-of-the-art language technologies for South Slavic languages and dialects, and managed an infrastructure dedicated to the continuous collection of massive web corpora for these languages.

Her work is published in international journals and conferences related to the fields of computational linguistics and natural language processing, including the LREC, COLING, and EACL conferences. As an author or co-author of over 70 resources, Taja Kuzman Pungeršek is one of the main contributors of text datasets and language technologies for Slovenian and other European languages within the CLARIN.SI repository.