machine learning & knowledge extraction

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

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

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

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 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.

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 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 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.

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.

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

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.

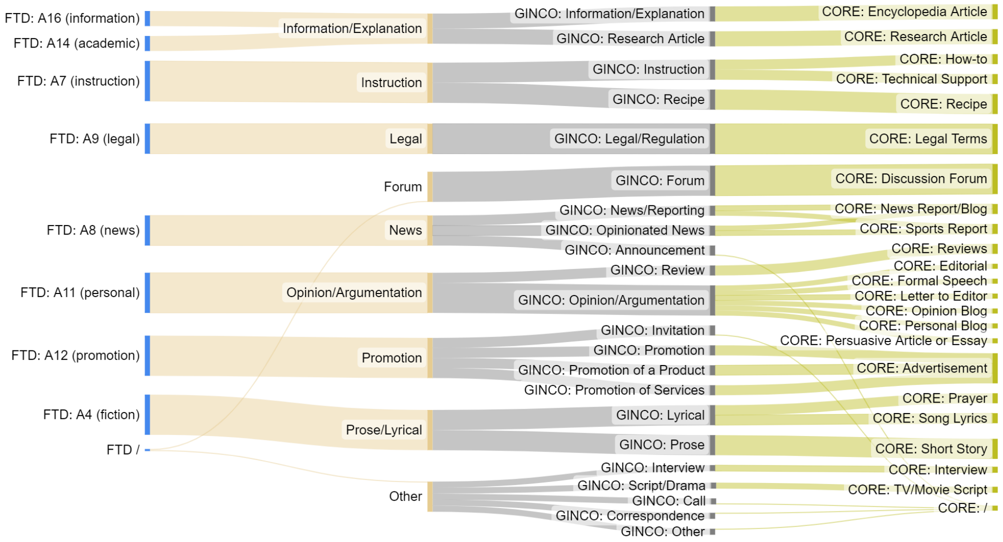

Figure 1. A diagram showing how labels from the FTD [28] (first column), GINCO [19] (third column), and CORE [17] (fourth column) schemata are mapped to the new joint X-GENRE schema (second column).

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.

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 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.

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.

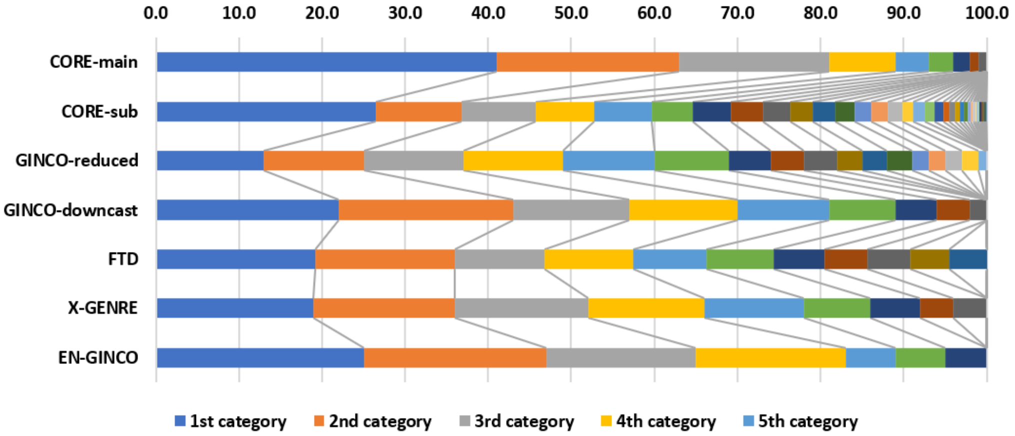

Figure 2. The comparison of category distributions in genre datasets in terms of percentages of instances belonging to each category, from the most frequent category (1st category) to the least frequent one.

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 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.

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 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, 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 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 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.

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.

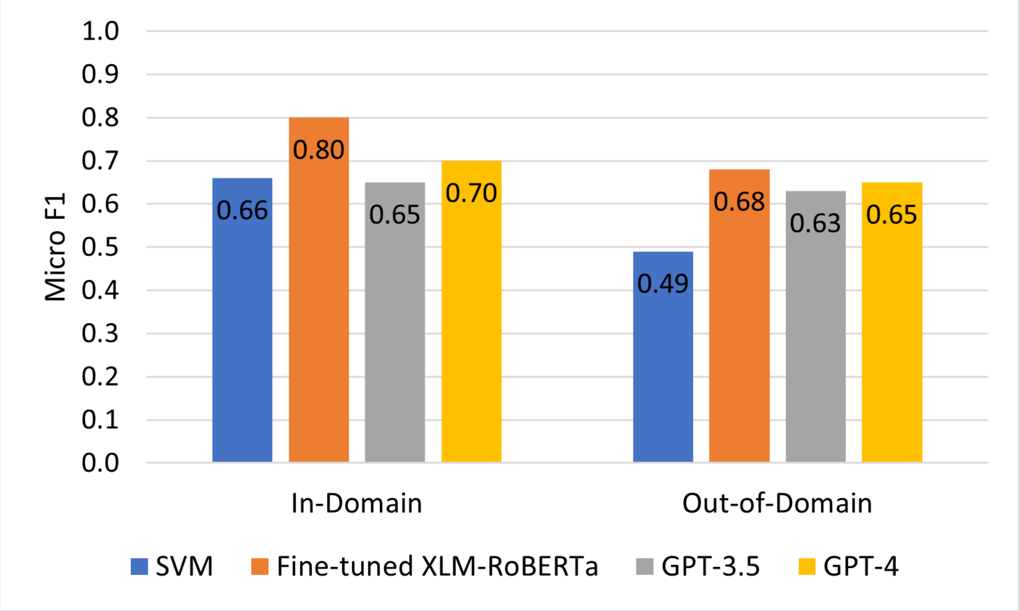

Figure 3. Overview of the performance of the SVM and fine-tuned XLM-RoBERTa model, which were trained on the training split of the X-GENRE dataset and evaluated in in-domain (X-GENRE test split) and out-of-domain (EN-GINCO test dataset) scenarios, and the performance of the GPT-3.5 and GPT-4 models, 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 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.

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

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]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.