KNOWLEDGE-GRAPH-INFORMED FAKE NEWS CLASSIFICATION VIA HETEROGENEOUS REPRESENTATION ENSEMBLES

Boshko Koloski

Master Thesis
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Assist. Prof. Dr. Senja Pollak, Jožef Stefan Institute, Ljubljana, Slovenia
Co-Supervisor: Dr. Blaž Škrlj, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Marko Robnik-Šikonja, Chair, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
Assist. Prof. Dr. Slavko Žitnik, Member, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
Assist. Prof. Dr. Senja Pollak, Member, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Boshko Koloski

KNOWLEDGE-GRAPH-INFORMED FAKE NEWS CLASSIFICATION VIA HETEROGENEOUS REPRESENTATION ENSEMBLES

Master Thesis

UPORABA GRAFOV ZNANJA PRI KLASIFIKACIJI LAŽNIH NOVIC Z ANSAMBLI HETEROGENIH REPREZENTACIJ

Magistrsko delo

Supervisor: Assist. Prof. Dr. Senja Pollak

Co-Supervisor: Dr. Blaž Škrlj

Ljubljana, Slovenia, July 2022

To my family.
Morior invictus.

Acknowledgments

I would like to express my gratitude to my thesis supervisors Assist. Prof. Dr. Senja Pollak. and Dr. Blaž Škrlj The advice, guidance and constant support throughout the duration of the research and my studies helped me grow both as a scientist and as a person.

I would like to thank Nada Lavrač for providing helpful insights and tips for improving my scientific journey.

I would like to thank Andraž Pelicon and Dr. Matej Martinc for bearing with me on a weekly basis and for the long hours of debates and constant improvement tips and tricks.

I would like to thank Mili Bauer for making my life easier by providing solutions to all of my bureaucratic endeavours.

I would like to express my appreciation for the financial support from the H2020 EMBEDDIA project (No. 825153), the Slovenian Research Agency for the research core funding for the programme Knowledge Technologies (No. P2-0103) and the project CANDAS (No. J6-2581) for funding my research and making my life so much easier.

I would like to thank my friends Marko, Ferdi, Gojko, Gorjan, Tomche, Ilir, Ljupche, Ema, Stefani, Jovan, Mihajlo (random order) for listening to my oddly specific scientific talks and for brightening up my free time.

I would like to thank Tea for bearing with me throughout the scientific journey, for the random cool facts and for just being here.

Finally, I must express my profound gratitude to my sister Nadica and my parents Levko and Violeta, for their unconditional support and love thoughtout the studies.

Abstract

Increasing amounts of freely available data both in textual and relational form offers exploration of richer document representations, and their potential for improving the performance and robustness of models. An emerging problem in the modern era, with a huge amount of information posted daily, is fake news detection — many easily available pieces of information are not necessarily factually correct, and can lead to wrong conclusions or are used for manipulation. The family of fake news problems can be split into three categories: fake news detection, fake news spreaders profiling and fact-checking. The first category addresses the problem of classifying a single document (a single news article or social-media post) in real or fake news. The second type of problems addresses the question if a particular author tends to spread fake news. The third and final category is based on verifying if a given statement is factually correct. In this study we focus only on the first and second category of fake news detection.

In recent years, there is a spike in the gathering and the curation of factual knowledge structured in knowledge bases. Researchers proposed various algorithms to convert the knowledge from the knowledge graphs to dense numeric representations. The algorithms are successful in capturing various implicit relations amongst concepts appearing in the knowledge bases.

In this thesis, we explore how different document representations, ranging from simple symbolic bag-of-words to contextual, neural language model-based ones, can be used for efficient fake news identification. One of the key contributions is a set of novel document representation learning methods based solely on knowledge graphs, i.e. extensive collections of subject-predicate-object triplets. We evaluate our method on four standard fake news detection data sets (one fake news spreaders profiling and three fake news classification data sets). We demonstrate that knowledge graph-based representations already alone achieve competitive performance to contemporary representation learners. Furthermore, when combined with contextual and non-contextual representations into heterogeneous ensembles of representations, knowledge graph-based document representations can contribute to achieving state-of-the-art performance. To our knowledge, this is the first larger-scale evaluation of how knowledge graph-based representations can be systematically incorporated into the process of fake news classification.

Povzetek

Vse večje količine prosto dostopnih podatkov v besedilni in relacijski obliki omogočajo raziskovanje bogatejših predstavitev dokumentov in njihov potencial za izboljšanje zmogljivosti in robustnosti modelov. Pojavljajoča se težava v današnji dobi, kjer je dnevno objavljena velika količina informacij, je odkrivanje lažnih novic — številni zlahka dostopni deli informacij niso nujno dejansko pravilni in lahko vodijo do napačnih sklepov ali se uporabljajo za manipulacijo. Naloge odkrivanja lažnih novic lahko razdelimo v tri kategorije: odkrivanje lažnih novic, profiliranje avtorjev, ki širijo lažne novice, in preverjanje dejstev. Prva kategorija naslavlja problem klasifikacije posameznega dokumenta (novičarskega članka ali objave na družbenem omrežju) v razred resničnih ali lažnih novic. Druga vrsta problemov obravnava vprašanje, ali določen avtor širi lažne novice. Tretja kategorija pa temelji na preverjanju dejanske resničnosti izjav. V tej študiji se osredotočamo le na prvo in drugo kategorijo odkrivanja lažnih novic.

V zadnjih letih je prišlo do porasta zbiranja in urejanja znanja v bazah znanja. Raziskovalci so razvili različne algoritme za pretvorbo informacij iz grafov znanja v goste številčne predstavitve. Algoritmi so uspešni pri zajemanju različnih implicitnih relacij med koncepti iz baz znanj.

V magistrskem delu raziskujemo, kako je mogoče različne predstavitve dokumentov, od preproste vreče besed do kontekstualnih, nevronskih jezikovnih modelov, uporabiti za učinkovito prepoznavanje lažnih novic. Eden ključnih prispevkov je niz novih učnih metod za reprezentacijo dokumentov, ki temeljijo izključno na grafih znanja, torej na obsežnih strukturiranih zbirkah trojic v obliki subjekt-predikat-objekt.

Našo metodo ocenimo na štirih standardnih zbirkah za razpoznavanje lažnih novic (ena zbirka besedil vsebuje podatke za profiliranje avtorjev, ki širijo lažne novice, tri pa zajemajo označene množice lažnih novic). Pokažemo, da predstavitve, ki temeljijo na grafih znanja, že same dosegajo konkurenčno uspešnost napram standardnim metodam za učenje reprezentacij. Poleg tega pa lahko predstavitve dokumentov na podlagi grafov znanja uporabimo v kombinaciji z drugimi kontekstualnimi in nekontekstualnimi predstavitvami in na podlagi ansamblov heterogenih reprezentacij dosežemo odlične rezultate. Kolikor nam je znano, je to prva obsežnejša študija, ki preučuje, kako lahko reprezentacije, ki temeljijo na grafih znanja, sistematično vključimo v proces klasifikacije lažnih novic.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Identifying fake news is a crucial task in the modern era. Fake news can have devastating implications on society; the uncontrolled spread of fake news can, for example, impact the idea of democracy, with the ability to alter the course of elections by targeted information spreading (Allcott & Gentzkow, 2017). In the times of a global pandemic fake news can endanger the global health, for example by reporting that using bleach can stop the spread of coronavirus (Kadam & Atre, 2020; Pulido et al., 2020), or that vaccines are problematic for human health. In the era of information society, the increasing capability to create and spread news in various formats makes detection of fake news even more important.

For media companies' reputation it is crucial to avoid distributing **unreliable information**. With the ever-increasing number of users and potential fake news spreaders, relying only on manual analysis is becoming unmanageable given the number of posts a single person can curate on a daily basis. Therefore, the need for *automated detection* of fake news is more important than ever, making it also a very relevant and attractive research task.

Several tasks have been formally proposed in the context of combating disinformation via automated methods. The basic task of *fake news detection* focuses on deciding for a single post (a short textual document) whether it is genuine or fake. Some variants of this problem go beyond classifying a document as real or fake and introduce multiple levels of truth (how true or fake is a particular news item). Instances of this type of problem can be domain invariant such as the Liar dataset (W. Y. Wang, 2017) acquired from Politico that contained news about various events, or they can be domain specific, such as the problem of *detection of COVID-19 fake news* (Patwa et al., 2020). The next task in the series concerns *profiling of a particular author* in a social medium with respect to their tendency to spread fake news. Formally, the inputs to this task are a list of documents and a label whether the author spreads fake news or not. Rangel et al. (2020) have proposed the problem of *identifying fake news spreaders on Twitter* in both English and Spanish. An extended variant of this problem takes into account both the given network of users associated with a given author and the network of users who have interacted with a given social media post. One such example is the work by Gupta and Potika (2021), where the analysis of Twitter fake news during the COVID-19 pandemic is considered at both the network and document levels. Graph-aware models are being used for this type of problems due to the social-network structure of this problem, where for a given user the corresponding sub-network of the social media is provided (Chandra et al., 2020; Hamid et al., 2020; Monti et al., 2019; Nguyen et al., 2020).

By being able to process large collections of labeled and unlabeled textual inputs, machine learning approaches are becoming a viable solution to (semi-)automatic fake news detection and credibility assessment (Shu et al., 2017). One of the key problems, how-

ever, concerns the representation of such data in a form suitable for learning. Substantial advancements were made in this direction in the last years, ranging from large-scale curated knowledge graphs to contextual language models capable of differentiating subtle differences between a multitude of texts (Shu et al., 2020). This thesis explores how such technologies can be used to aid and prevent spreading of problematic content, at scale. We focus on the *fake news detection* and *fake news spreaders identification* tasks.

With the advancements in the field of machine learning and natural language processing, various different computer-understandable representations of texts have been proposed. While the recent work has shown that leveraging background knowledge can improve document classification (Ostendorff et al., 2019), this path has not yet been sufficiently explored for fake news identification. The main contributions of this thesis are:

1. We explore how additional background knowledge in the form of **knowledge graphs**, constructed from freely available knowledge bases, can be exploited to enrich various contextual and non-contextual document representations.

2. We conducted extensive experiments where we systematically studied the effect of five document and six different knowledge graph-based representations on the model performance.

3. We propose a feature-ranking-based *post-hoc* analysis capable of pinpointing the key types of representation, relevant for a given classification problem.

4. The explanations of the best-performing model are inspected and linked to the existing domain knowledge.

The results of this thesis have been published in a number of papers. First, an LSA approach has been proposed to identify fake news spreaders (Koloski, Pollak, et al., 2020). Next, we proposed an approach for fake news identification using representation ensembles (Koloski et al., 2021). The approach was then extended by including novel knowledge-graph-based embeddings, which also represents the majority of the work covered in this master thesis and was published in the *Neurocomputing* journal paper (Koloski et al., 2022). In addition, our knowledge-graph representations can improve several other tasks, as we showed in Tavchioski et al. (2022) and in Zosa et al. (2022).

The remaining work is structured as follows. In Chapter 2, we present the relevant related work, followed by the text and graph representations used in our study, in Chapter 3 we present the proposed method, followed by the evaluation in Chapter 4. We discuss the obtained results in Sections 5.1 and 5.2 and finish with the concluding remarks in Chapter 6.

Chapter 2

Related Work

We next discuss the considered classification task and the existing body of literature related to identification/detection of fake news. The fake news text classification task is defined as follows: given a text and a set of possible classes (e.g., fake and real) to which a text can belong, an algorithm is tasked with predicting the correct class label assigned to the text. Most frequently, fake news text classification refers to classification of data based on social media.

The early proposed solutions to this problem used hand-crafted features of the authors (instances) such as word and character frequencies (Potthast et al., 2018). A more advanced incorporation of the stylometric features was introduced by Buda and Bolonyai (2020) where apart from them, n-gram-based features were introduced and fused together into an ensemble of models. The ensemble-based modeling was also explored by Hörtenhuemer and Zangerle (2020), where authors build models on multi-aspect features (such as sentiment, named entities, readability score, emotional analysis, word and character n-grams), and finally combined into an ensemble. The introduction of features using transfer-learning-based emotional analysis showcased to be a good baseline in the work of Murrieta Bello et al. (2020). Multilingual identification of fake news based on latent space of n-grams for Spanish and English fake news was performed by Koloski, Pollak, et al. (2020). Many of the contemporary machine learning approaches are based on deep neural-network models (Glazkova et al., 2020).

Currently, the transformer architecture (Vaswani et al., 2017) is commonly adopted for various down-stream learning tasks. The winning solution to the COVID-19 Fake News Detection task (Patwa et al., 2020) utilized fine-tuned BERT model that considered Twitter data scraped in the COVID-19 period (Müller et al., 2020) - January 12 to April 16, 2020 (Glazkova et al., 2020). Kaliyar et al. (2021) proposed FakeBERT - model built on top of the BERT model and a single-layer deep convolutional network. The authors claim that this approach improved the handling of ambiguity as one of the major challenges in natural language understanding and consequently improved the detection of fake news. Other solutions exploited the recent advancements in the field of Graph Neural Networks and their applications in these classification tasks (J. Zhang et al., 2020). However, for some tasks best preforming models are based on traditional n-gram feature crafted representations and a linear learners like SVM, learned on top of them (Buda & Bolonyai, 2020).

Interestingly, the stylometry-based approaches were shown to be a potential threat for the automatic detection of fake news (Schuster et al., 2020) . The reason for this is that machines are able to generate consistent writings regardless of the topic, while humans tend to be biased and make some inconsistent errors while writing different topics. Additionally researchers explored how the traditional machine learning algorithms perform on such

tasks given a single representation (Gilda, 2017). The popularity of deep learning and the successes of Convolutional and Recurrent Neural Networks motivated the development of models following these architectures for the tasks of headline and text matching of an article (Umer et al., 2020).

Lu and Li (2020) proposed a solution to a more realistic scenario for detecting fake news on social media platforms which incorporated the use of graph co-attention networks on the information about the news, but also about the authors and spread of the news. A multi-modal approach studied the correlation between text and images via deep-neural networks for improved combating of fake news infodemic in the work of Zeng et al. (2021). Leveraging of the social context for a given article such as community information, the authors' profiles on social media and the content of the article was explored via tensor-decomposition-based deep neural network in the work of Kaliyar et al. (2021). The trend of combining different classifiers into ensembles was studied also on transformer-based architectures (Balouchzahi et al., 2021). Jiang et al. (2021) studied the stacking of various traditional and contemporary model architectures based on both contextual and non-contextual features.

Despite the fact that the neural network-based approaches outperform other approaches on many tasks, they are not directly **interpretable** (Lipton, 2018). On the other hand, more traditional machine learning methods such as symbolic and linear models are easier to interpret and reason with, despite being outperformed by contemporary deep-learning methods. To incorporate both viewpoints, a significant amount of research has been devoted to the field of **neuro-symbolic computing**, which aims to bring the robustness of neural networks and the interpretability of symbolic approaches together.

Knowledge graphs have recently shown to be a performance-boosting aid in various domains. An improvement on the task of reasoning and question answering was achieved by the incorporation of knowledge graphs into language models (Yasunaga et al., 2021). Enrichment of the contextualized language models for the domain of biomedicine with domain-specific knowledge graph improved the results in the task of information extraction (Fei et al., 2020). In a more recent study (Moiseev et al., 2022), a method was proposed to infuse structured knowledge data to the large language models. The new knowledge-infused models showed superior performance to the standalone language models on multiple tasks. An improvement of multi-event prediction was achieved with knowledge-aware networks (Song et al., 2021). Researchers have also explored the benefits of deriving knowledge-aware representations. For example, a recent approach explored document representation enrichment with symbolic knowledge (Z. Wang et al., 2014). In their approach, the authors tried enriching a two-part model: a text-based model consisting of statistical information about text and a knowledge model based on entities appearing in both the KG and the text. Further, Ostendorff et al., 2019 explored a similar idea considering learning separate embeddings of knowledge graphs and texts, and later fusing them together into a single representation. An extension to the work of Ostendorff et al. (2019) was preformed by Koloski, Škrlj, et al. (2020), where a promising improvement of the joint representations has been observed. This approach showed potentially useful results, improving the performance over solely text-based models. Hu et al. (2021) introduced knowledge graphs to the task of fake news and achieved state-of-the-art results on one data set. In their work, first a document graph is built composed of topics and named entities, next a graph attention network is utilized for learning of the topic-enriched news and finally knowledge-based entity representation is derived via entity comparison network. In the same manner, Dun et al. (2021) proposed a knowledge-aware attention network composed of deriving aggregated contextual embedding of entities appearing in a document. For the entities present in the document and the knowledge graph, they search the concepts that are one-hop away in

the knowledge graph and embed them with a word2vec (Mikolov et al., 2013) method. Finally, they fuse additional document representations with the derived ones in a deep-neural network with an attention mechanism.

Different approaches achieve state-of-the-art results when considering various tasks related to fake news detection. However, individual representations of documents suitable for solving a given problem are mostly problem-dependent, motivating us to explore *representation ensembles*, which potentially entail different aspects of the represented text, and thus generalize better.

While fake news classification is widespread, most practically useful systems consider fake news classification as a matter of trust in the source as the text itself. As one can write a news in a convincing style, even if it contains misinformation, fake news detection problem as a language-only problem has limitations. However, when no information is given regarding the source, one needs to rely on language-only document representation. This viewpoint of the problem calls for different representation structure to be included while building fake news detection systems. knowledge graphs as ground truth knowledge bases serve as a complement to text-only representations and allow for greater generalisation and as such can be included in the detection of fake news.

Chapter 3

Methodology

In this chapter, we explain the proposed knowledge-based representation enrichment method. First we define the relevant document representations, followed by concept extraction and knowledge graph (KG) embedding. Finally, we present the proposed combination of the constructed feature spaces. Schematic overview of the proposed methodology is shown in Figure 3.1.

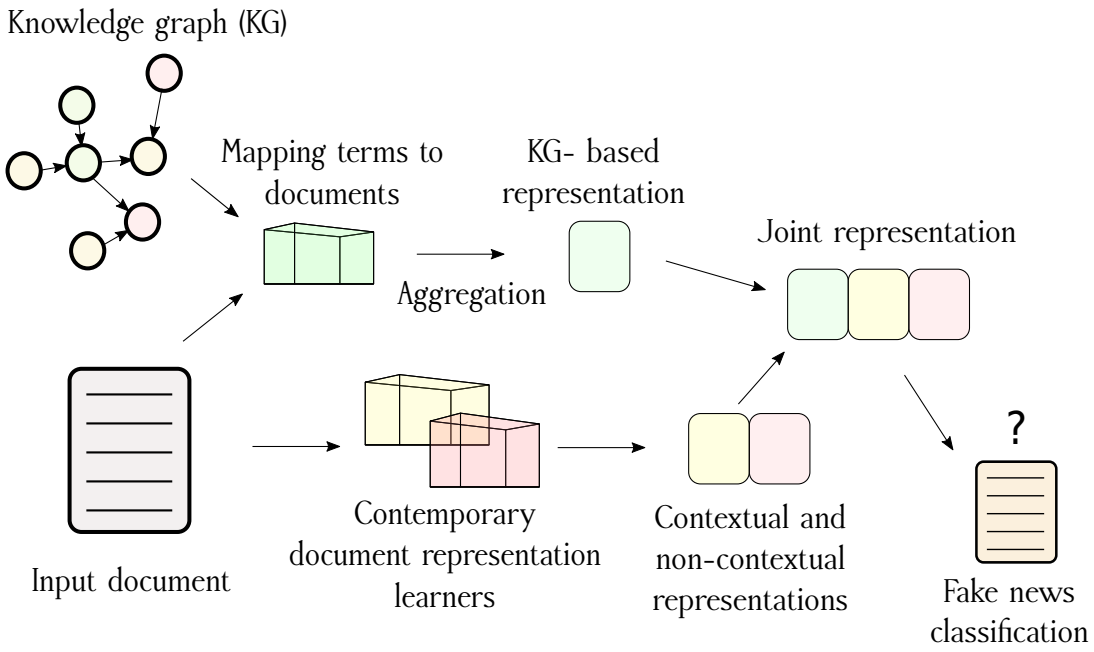


Figure 3.1: Schematic overview of the proposed methodology. Both knowledge graph-based features and contextual and non-contextual document features are constructed, and used simultaneously for the task of text classification.

We begin by describing the bottom part of the scheme (yellow and red boxes), followed by the discussion of KG-based representations (green box). Finally, we discuss how the representations are combined ("Joint representation") and learned from (final step of the scheme).

3.1 Existing Document Representations Considered

Various document representations capture different patterns across the documents. For the text-based representations we focused on exploring and exploiting the methods we already developed in our submission to the COVID-19 fake news detection task (Koloski et al., 2021). We next discuss the document representations considered, which are the main contributions of this thesis (and the paper (Koloski et al., 2022)).

Hand-Crafted features. We use stylometric features inspired by early work in authorship attribution (Potthast et al., 2018). We focused on word-level and character-level statistical features.

- **Word-based features**. The word-based features included maximum and minimum word length in a document, average word length, and standard deviation of the word length in document. Additionally, we counted the number of words beginning with an upper and the number of words beginning with a lower case.
- **Character-based features** The character-based features consisted of the counts of digits, letters, spaces, punctuation, hashtags and each vowel, respectively.

Hence, the final statistical representation has 10 features.

Latent Semantic Analysis. Similarly to the Koloski, Pollak, et al. (2020) solution to the PAN 2020 shared task on Profiling Fake News Spreaders on Twitter (Rangel et al., 2020) we applied the low dimensional space estimation technique. First, we preprocessed the data by lower-casing the document content and removing the hashtags, punctuation and stop words. From the cleaned text, we generated the POS-tags using the NLTK library (Loper & Bird, 2002). Next, we used the prepared data for feature construction. For the feature construction we used the technique used by Martinc et al. (2018) which iteratively weights and chooses the best n-grams. We used two types of n-grams: Word-based: n-grams of size 1 and 2 and Character-based: n-grams of sizes 1, 2 and 3. We generated word and character n-grams and used TF-IDF for their weighting. We performed SVD (Halko et al., 2009) of the TF-IDF matrix, where we only selected the m most-frequent n-grams from word and character n-grams. With the last step we obtained the LSA representation of the documents. For each of our tasks, our final representation consists of 2,500 word and 2,500 character features (i.e. 5,000 features in total) reduced to 512 dimensions with the SVD. The original paper considers this combination of features and dimensions as most suitable for representation of short texts.

Contextual features. For capturing contextual features we utilize embedding methods that rely on the transformer architecture (Vaswani et al., 2017), including:

- DistilBert (Sanh et al., 2019) *distilbert-base-nli-mean-tokens* - d = 768 dimensions
- RoBERTa (Liu et al., 2019) - *roberta-large-nli-stsb-mean-tokens* - d = 768 dimensions
- XLM (Conneau & Lample, 2019) - *xlm-r-large-en-ko-nli-ststb* - d = 768 dimensions

First, we applied the same preprocessing as in *Latent Semantic Analysis* representations. After we obtained the preprocessed texts we embedded every text with a given transformer model and obtained the contextual vector representation. As the transformer models work with a limited number of tokens, the obtained representations

were 512-dimensional, as this was the property of the used pre-trained models. This did not represent a drawback since most of the data available was shorter than this maximum length. The contextual representations were obtained via pooling-based aggregation of intermediate layers (Reimers & Gurevych, 2019).

3.2 Knowledge Graph-Based Document Representations

We continue the discussion by presenting the key novelty of this thesis: document representations based solely on the existing background knowledge. To be easily accessible, human knowledge can be stored as a collection of facts in knowledge bases (KB). The most common way of representing human knowledge is by connecting two entities with a given relationship that relates them. Formally, a knowledge graph can be understood as a directed multigraph, where both nodes and links (relations) are typed. A concept can be an abstract idea such as a thought, a real-world entity such as a person, e.g., Donald Trump, or an object - a vaccine, and so on. An example fact is the following: Ljubljana (entity) is the capital (relation) of Slovenia (entity), the factual representation of it is *(Ljubljana,capital,Slovenia)*. Relations have various properties, for example the relation *sibling* that captures the symmetry-property - if (Ann,siblingOf,Bob) then (Bob,siblingOf,Ann), or antisymmetric relation fatherOf (Bob,fatherOf,John) then the reverse does not hold (John,fatherOf,Bob).

In order to learn and extract patterns from facts, the computers need to represent them in a useful manner. To obtain the representations we use six knowledge graph embedding techniques: TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), QuatE (S. Zhang et al., 2019), ComplEx (Trouillon et al., 2016), DistMult (Yang et al., 2015) and SimplE (Kazemi & Poole, 2018). The goal of a knowledge graph embedding method is to obtain numerical representation of the KG, or in the case of this thesis, its entities. The considered KG embedding methods also aim to preserve relationships between entities. The aforementioned methods and the corresponding relationships they preserve are listed in Table 3.1. It can be observed that RotatE is the only method capable of modeling all five relations due to its specific modeling of relations as *rotations* in a complex numeric space. Even though other methods are theoretically not as expressive, this does not indicate their uselessness when considering construction of document representations. For example, if transitivity is crucial for a given data set, and two methods, which theoretically both model this relation, capture it to a different extent, even simpler (and faster) methods such as TransE can perform well.

We propose a novel method for combining background knowledge in the form of a knowledge graph KG about concepts C appearing in the data D. To transform the documents in numerical spaces we utilize the techniques described previously. For each technique we

Table 3.1: Relations captured by specific knowledge graph embedding from the GraphVite knowledge graph suite (Zhu et al., 2019).

Name	Symmetry	Anti-symmetry	Inversion	Transitivity	Composition
TransE (Bordes et al., 2013)	**x**	**x**	✓	✓	**x**
DistMult (Yang et al., 2015)	✓	**x**	**x**	**x**	**x**
ComplEx (Trouillon et al., 2016)	✓	✓	✓	✓	**x**
RotatE (Sun et al., 2019)	✓	✓	✓	✓	✓
QuatE (S. Zhang et al., 2019)	✓	✓	✓	✓	**x**
SimplE (Kazemi & Poole, 2018)	✓	✓	✓	✓	**x**

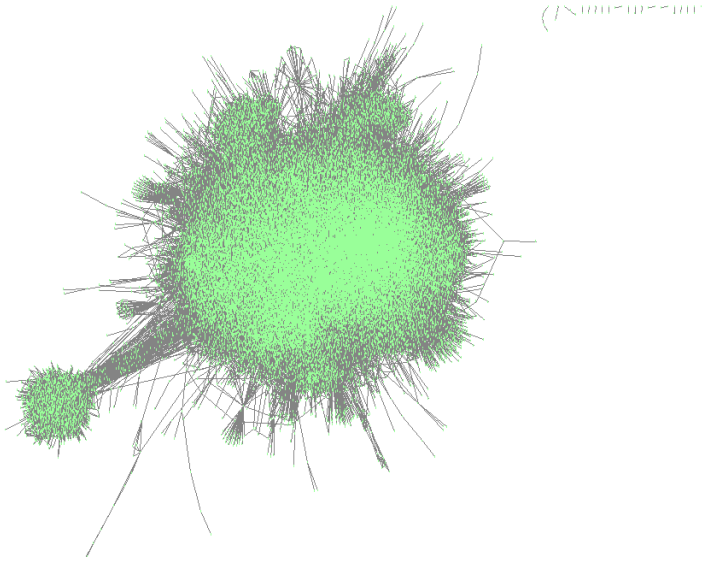


Figure 3.2: The WikiData5m knowledge graph - the ≈100,000 most connected nodes. It can be observed that multiple smaller structures co-exist as part of the global, well connected- structure.

learn the space separately and later combine them in order to obtain the higher dimensional vector spaces useful for solving a given classification task.

For representing a given document, the proposed approach can consider the document text or also account for additional metadata provided for the document (e.g. the author of the text, their affiliation, who is the document talking about etc.). In the first case, we identify which concept embeddings map to a given piece of text, while in the second scenario, we also embed the available metadata and jointly construct the final representation. In this study, we use the WikiData5m knowledge graph (Vrandečić & Krötzsch, 2014) (Figure 3.2 for the visualisation of the 100,00 most connected nodes in this KG). The most central nodes include terms such as 'encyclopedia' and 'united state'.

The GraphVite library [1] (Zhu et al., 2019) incorporates approaches that map aliases of concepts and entities into their corresponding embeddings. To extract the concepts from the documents we first preprocess the documents with the following pipeline: punctuation removal; stopword removal for words appearing in the NLTK's english stopword list; lemmatization via the *NLTK's WordNetLemmatizer* tool.

In the obtained texts, we search for concepts (token sets) consisting of uni-grams, bi-grams and tri-grams, appearing in the knowledge graph. The concepts are identified via exact string alignment. With this step we obtained a collection of candidate concepts C_d for each document d.

From the obtained candidate concepts that map to each document, we developed three different strategies for constructing the final representation. Let $\boldsymbol{e}^i$ represent the i-th dimension of the embedding of a given concept. Let $\bigoplus$ represent the element-wise summation (i-th dimensions are summed). We consider the following aggregation. We considered using all the concepts with equal weights and obtained final concept as the average of the concept embeddings:

$$\text{AGG-AVERAGE}(C_d) = \frac{1}{|C_d|} \bigoplus_{c \in C_d} \boldsymbol{e}_c.$$

[1] https://github.com/DeepGraphLearning/graphvite

Table 3.2: Summary table of the textual and KG representations used in this paper.

Name	Type	Description	Dimension
Stylomteric	text	Statistical features capturing style of an author.	10
LSA	text	N-gram based representations built on chars and words reduced to lower dimension via SVD.	512
DistilBert	text	Contextual - transformer based representation learned via sentence-transformers.	768
XLM	text	Contextual - transformer based representation learned via sentence-transformers.	768
RoBERTa	text	Contextual - - transformer based representation learned via sentence-transformers.	768
TransE	KG	KG embedding capturing inversion, transitivity and composition property.	512
DistMult	KG	KG embedding capturing symmetry property.	512
ComplEx	KG	KG embedding capturing symmetry, anti-symmetry, inversion and transitivity property.	512
RotatE	KG	KG embedding captures inversion, transitivity and composition property.	512
QuatE	KG	KG embedding capturing symmetry, anti-symmetry, inversion, transitivity and composition property.	512
SimplE	KG	KG embedding capturing symmetry, anti-symmetry, inversion and transitivity property.	512

The considered aggregation scheme, albeit one of the simpler ones, already offered document representations competitive to many existing mainstream approaches. The key parameter for such representations was embedding dimension, which was in this thesis set to 512.

3.3 Construction of the Final Representation

Having presented how document representations can be obtained from knowledge graphs, we next present an overview of the considered document representations used for subsequent learning, followed by the considered representation combinations. The overview is given in Table 3.2. Overall, 11 different document representations were considered. Six of them are based on knowledge graph-based embedding methods. The remaining methods either consider contextual document representations (RoBERTa, XLM, DistilBert), or non-contextual representations (LSA and stylometric). The considered representations entail multiple different sources of relevant information, spanning from single character-based features to the background knowledge-based ones.

For exploiting the capability of the multi-modal representations we consider three different scenarios to compare and study the potential of the representations:

LM - we concatenate the representations from Section 3.1 - handcrafted statistical features, Latent Semantic Analysis features, and contextual representations - XLM, RoBERTa and DistilBERT.

KG - we concatenate the aggregated concept embeddings for each KG embedding method from Section 3.2 - TransE TransE, SimplE, ComplEx, QuatE, RotatE and DistMult. We agreggate the concepts with the AGG-AVERAGE strategy.

Merged - we concatenate the obtained language-model and knowledge graph representations. As previously mentioned we encounter two different scenarios for KG enriched representations:

- LM+KG - we combine the induced KG representations with the methods explained in Section 3.2.
- LM+KG+KG-ENTITY - we combine the document representations, induced KG representations from the KG and the metadata KG representation if it is available. To better understand how the metadata are used (if present), consider the following example. Consider a document, for the author of which we know also the following information: *speaker = Dwayne Bohac, job = State representative, subject = abortion, country = Texas, party affiliation = republican.*

The values of such metadata fields (e.g., job) are considered as any other token, and checked for their presence in the collection of knowledge graph-based entity embeddings. Should the token have a corresponding embedding, it is considered for constructing the KG-ENTITY representation of a given document. For the data sets where the metadata is present, it is present for all instances (documents). If there is no mapping between a given collection of metadata and the set of entity embeddings, empty (zero-only) representation is considered.

Having discussed how the constructed document representation can be combined systematically, we next present the final part needed for classification – the representation ensemble construction.

3.4 Classification Models Considered

We next present the different neural and non-neural learners, which consider the constructed representations discussed in the previous section.

Representation stacking with linear models. The first approach to utilize the obtained representations was via linear models that took the stacked representations and learned a classifier on them. We considered using a LogisticRegression learner and a StochasticGradientDescent-based learner that were optimized via either a *log* Vovk (2015) or *hinge* Gentile and Warmuth (1998) loss function. We applied the learners on the three different representations scenarios.

Representation stacking with neural networks. Since we have various representations of text and the concepts appearing in the data we propose an intermediate joint representation to be learnt with a neural network. For this purpose, we propose stacking the inputs in a heterogeneous representation and learning intermediate representations from them with a neural network architecture. The schema of our proposed neural network approach is represented in Figure 3.3. We tested three different neural networks for learning this task.

The proposed architecture consists of two main blocks: the input block and the hidden layers-containing block. The input block takes the various representations as parameters and produces a single concatenated representation which is normalized later. The hidden layer block is the learnable part of the architecture, the input to this block are the normalized representations and the number of the intermediate layers as well as their dimension. We evaluate three variants of the aforementioned architecture:

[SNN] Shallow neural network. In this neural network we use a single hidden layer to learn the joint representation.

[5Net] Five-hidden-layer neural network. The original approach that we proposed to solve the COVID-19 Fake News Detection problem featured a five-layer neural network to learn the intermediate representation (Koloski et al., 2021). We alter the original network with the KG representations for the input layer.

[LNN] Log(2) scaled neural network. Deeper neural networks in some cases appear to be more suitable for some representation learning tasks. To exploit this hypothesis we propose a deeper neural network - with a domino-based decay. For n intermediate layers we propose the first intermediate layer to consist of 2^n neurons, the second to be with 2^{n-1} ... and the n_0-th to be the activation layer with the number of unique outputs.

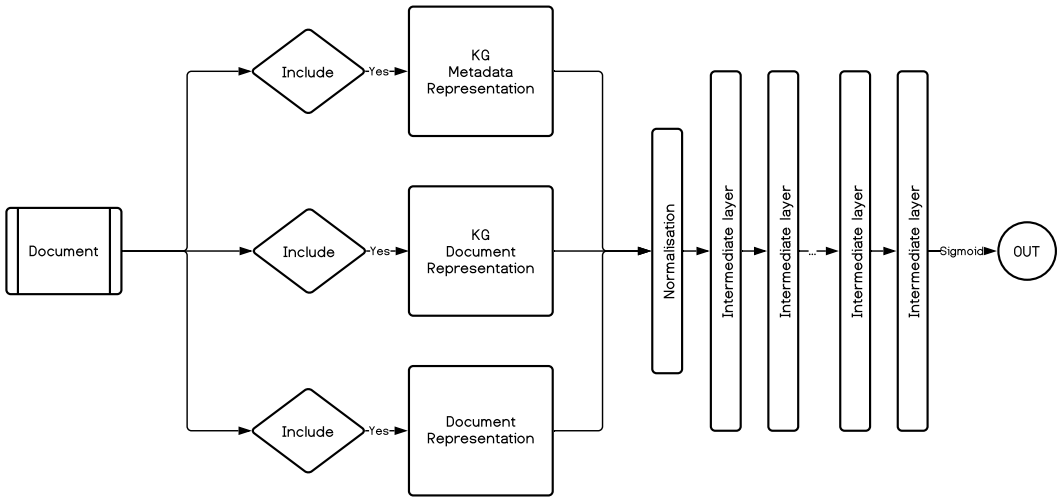


Figure 3.3: Neural network architecture for learning the joint intermediate representations. The *Include* decision block implies that some of the representations can be optionally excluded from the learning. The *Normalizaiton layer* normalizes the input to prevent skewed gradients. The number of the intermediate layers and the dimensions are of varying sizes and are part of the model's input. The final output presents the model's probability for a given label to be considered for the given document.

Chapter 4

Empirical Evaluation

In this chapter, we first describe four data sets which we use for benchmarking of our method. Next we discuss the empirical evaluation of the proposed method, focusing on the problem of fake news detection.

4.1 Data Sets

In order to evaluate our method we use four different fake news problems. We consider a fake news spreaders identification problem, two binary fake news detection problems and a multi-label fake news detection problem. We next discuss the data sets related to each problem considered.

COVID-19 Fake News detection data set (Patwa et al., (2021, 2020)) is a collection of social media posts from various social media platforms: Twitter, Facebook, and YouTube. The data contains COVID-19 related posts, comments and news, labeled as *real* or *fake*, depending on their truthfulness. Originally the data is split in to three different sets: train, validation and test.

Liar, Liar Pants on Fire (W. Y. Wang, 2017) represents a subset of PolitiFact's collection of news that are labeled with different categories based on their truthfulness. PolitiFact represents a fact verification organization that collects and rates the truthfulness of claims by officials and organizations. This problem is multi-label classification-based on six degrees of fake news. For each news article, an additional metadata is provided consisting of speaker, controversial statement, US party to which the subject belongs, what is the subject of the statement and the occupation of the speaker.

Profiling Fake News Spreaders is an author profiling task that was organized under the PAN2020 workshop (Rangel et al., 2020). In author profiling tasks, the goal is to decide if an author is a spreader of fake news or not, based on a collection of posts the author published. The problem is proposed in two languages, English and Spanish. For each author, 100 tweets are given, which we concatenate as a single document representing that author.

FNID: FakeNewsNet (Amirkhani, 2020) is a data set containing news from the PolitiFact website. The task is binary classification with two different labels - real and fake. For each news article - fulltext, speaker and the controversial statement are given.

The data splits are summarised in Table 4.1.

Table 4.1: Distribution of samples per given label in the three splits: train, validation and test for all four data sets, respectively.

data set	Label	Train	Validation	Test
COVID-19	real	3360 (52%)	1120 (52%)	1120 (52%)
	fake	3060 (48%)	1020 (48%)	1020 (48%)
	all	6420 (100%)	2140 (100%)	2140 (100%)
PAN2020	real	135 (50%)	15 (50%)	100 (50%)
	fake	135 (50%)	15 (50%)	100 (50%)
	all	270 (100%)	30 (100%)	200 (100%)
FakeNewsNet	real	7591 (50.09%)	540 (51.03%)	1120 (60.34%)
	fake	7621 (49.91%)	518 (48.96%)	1020 (39.66%)
	all	15212 (100%)	1058 (100%)	1054 (100%)
LIAR	barely-true	1654 (16.15%)	237 (18.46%)	212 (16.73%)
	false	1995 (19.48%)	263 (20.48%)	249 (19.65%)
	half-true	2114 (20.64%)	248 (19.31%)	265 (20.92%)
	mostly-true	1962 (19.16%)	251 (19.55%)	241 (19.02%)
	pants-fire	839 (8.19%)	116 (9.03%)	92 (7.26%)
	true	1676 (16.37%)	169 (13.16%)	208 (16.42%)
	all	10240 (100%)	1284 (100%)	1267 (100%)

4.2 Document to Knowledge Graph Mapping

For each article we extract the uni-grams, bi-grams and tri-grams that also appear in the Wikidata5M KG. Additionally, for the *Liar* and the *FakeNewsNet* data sets we provided KG embedding based on the aggregated concept embedding from their metadata. In the case of the *Liar* data set, we present, the speaker, the party he represents, the country the speech is related with and the topic of their claim. In all evaluation experiments we use the AGG-AVERAGE aggregation of concepts.

4.3 Classification Setting

We use the train splits of each data set to learn the models, and use the validation data splits to select the best-performing model to be used for the final test set evaluation. For both the linear stacking and the neural stacking we define custom grids for hyperparameter optimization, explained in the following subsections.

Learning of linear models For each problem we first learn a baseline model from the given representation and a L2 regularized Linear Regression with the parameter $\lambda_2 \in \{0.1, 0.01, 0.001\}$. We also learned StochasticGradientDescent(SGD)-based linear learner, optimizing 'log' and 'hinge' functions with ElasticNet Zou and Hastie, 2005 regularization. To optimize the SGD learner we defined a custom grid. We opted for a parameter grid that would offer various tight and flexible penalizations of learners, to be able to adapt it to different problems. We defined the following hyper-parameter grid:

$l1\_ratio \in \{0.05, 0.25, 0.3, 0.6, 0.8, 0.95\}$,

$power\_t \in \{0.1, 0.5, 0.9\}$,

$alpha \in \{0.01, 0.001, 0.0001, 0.0005\}$.

Learning of neural models The optimization function for all of the neural models was the CrossEntropyLoss optimized with the Adam optimizer (Kingma & Ba, 2015). We

used the *SELU* - Scaled Exponential Linear Unit Klambauer et al. (2017) function as an activation function between the intermediate layers. For fine-tuning purposes we defined a custom grid consisting of the learning rate λ, the dropout rate p and the number of intermediate layers n (for each network separately). The search spaces of each parameter are:

Learning rate: $\lambda \in \{0.0001, 0.005, 0.001, 0.005, 0.01, 0.05, 0.1\}$.

Dropout rate: $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Intermediate layer parameters:

- SN $n \in \{32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$.
- 5Net fixed sizes as in Koloski et al. (2021).
- LNN $n \in 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16$ which produced n. intermediate layers of sizes $2^n, 2^{n-1}, 2^{n-2}, ..., 2^2, 2$. Note that in total, ten different architectures were tested.

We considered batches of size 32, and trained the model for a maximum of 1,000 epochs with an early stopping criterion - if the result did not improve for 10 successive epochs, we stopped the optimization.

4.4 Baselines

The proposed representation-learner combinations were trained and validated by using the same split structure as provided in a given shared task. Hence, we compared our approach to the state-of-the-art for each data set separately. As the performance metrics differ from data set to data set, we compare our approach with the state-of-the-art with regard to the metric that was selected by the shared task organizers. We use four different evaluation metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Here, *TP* denotes True Positive - the amount of predictions that the model sees as positive and are indeed positive, *TN* denotes True Negative - the amount of predictions that the model sees as negative and are indeed negative. Similarly, *FP* denotes the amount of instances that are originally negative but the model labels them as positive, while *FN* is the opposite case, the amount of instances the model sees as negative but are indeed positive.

Chapter 5

Evaluation

In this chapter, we evaluate the proposed document representation method in a quantitative manner (Section 5.2) and a qualitative manner (Section 5.1). We compare the proposed methodology with the state-of-the-art for the corresponding task in the quantitative results. We also perform qualitative analysis on the feature importance for a given methodology.

5.1 Quantitative Results

In this section, we evaluate and compare the quality of the representations obtained for each problem described in Chapter 4. For each task we report four metrics: *accuracy*, *F1-score*, *precision* and *recall*.

5.1.1 Task 1: LIAR

The best-performing model on the validation set was a **[SNN]** shallow neural network with 128 neurons in the intermediate layer, a learning rate of 0.0003, batch size of 32, and a dropout rate of 0.2. The combination of the textual and KG representations improved significantly over the baseline models. The best-performing representations were constructed from the language model and the KG entities including the ones extracted from the metadata. The assembling of representations gradually improves the scores, with the combined representation being our top performing our model. The metadata-entity-based representation outperforms the induced representations by a margin of 2.42%, this is due to the captured relations between the entities from the metadata. The state-of-the-art model (Alhindi et al., 2018) is based on Bilinear Long Short Term Memory (BiLSTM) (Hochreiter & Schmidhuber, 1997) neural network that utilizes the Glove Pennington et al., 2014 embeddings and includes external information about a claim (such as extracted justification in conjunction with the claim). The evaluation of the task with respect to the models is shown in Table 5.1.

Table 5.1: Comparison of representations on the *Liar* data set without background knowledge (LM) with models incorporating text knowledge graph embeddings (KG) and metadata knowledge graph-embeddings (KG-ENTITY). LR in the representation column denotes the linear regression learner and SNN indicates the shallow neural network. The introduction of the factual knowledge constantly improved the performance of the model.

Representation	Accuracy	F1 - score	Precision	Recall
LR(LM)	0.2352	0.2356	0.2364	0.2352
LR(KG)	0.1996	0.1993	0.2004	0.1997
LR(LM + KG)	0.2384	0.2383	0.2383	0.2384
LR(KG-ENTITY)	0.2238	0.2383	0.2418	0.2415
LR(LM + KG-ENTITY)	0.2399	0.2402	0.2409	0.2399
LR(LM + KG + KG-ENTITY)	0.2333	0.2336	0.2332	0.2336
SNN(LM + KG + KG-ENTITY)	0.2675	**0.2672**	**0.2673**	**0.2676**
SOTA (literature) (Alhindi et al., 2018)	**0.3740**	x	x	x

5.1.2 Task 2: FakeNewsNet

The LNN was the best performing one for the *FakeNewsNet* problem with the n-parameter set to 12, a learning rate of 0.001, and a dropout of 0.7. The constructed KG representations outperformed both the LM representation by 1.99% and the KG-ENTITY representation by 2.19% in terms of accuracy and also outperformed them in terms of F1-score. The further combination of the metadata and the constructed KG features introduced significant improvement both with the linear stacking and the joint neural stacking, improving the baseline score by 1.23% for accuracy, 1.87% for F1-score and 3.31% recall for the linear stacking. The intermediate representations outscored every other representation by introducing 12.99% accuracy improvement, 13.32% improvement of F1-score and 26.70% gain in recall score. The proposed method improves the score over the current best performing model by a margin of 3.22%. The SOTA model for this task focuses on Natural Language Inference via a BiLSTM network built on top of contextual and non-contextual features. The evaluation of the task with respect to the models is shown in Table 5.2.

Table 5.2: Comparison of representations on the *FakeNewsNet* data set without background knowledge (LM) with models incorporating text knowledge graph embeddings (KG) and metadata knowledge graph-embeddings (KG-ENTITY). LR in the representation column denotes the linear regression learner and LNN indicates the use of the Log(2) neural network.

Representation	Accuracy	F1 - score	Precision	Recall
LR(LM)	0.7581	0.7560	0.9657	0.6210
LR(KG)	0.7780	0.7767	**0.9879**	0.6399
LR(LM+KG)	0.7676	0.7704	0.9536	0.6462
LR(KG-ENTITY)	0.7561	0.7512	0.9773	0.6100
LR(LM + KG-ENTITY)	0.7600	0.7602	0.9570	0.6305
LR(LM + KG + KG-ENTITY)	0.7704	0.7747	0.9498	0.6541
LNN(LM + KG + KG-ENTITY)	**0.8880**	**0.8892**	0.9011	**0.8880**
SOTA (literature) (Bidgoly et al., 2020)	0.8558	x	x	x

Table 5.3: Comparison of representations on the *PAN2020* data set without background knowledge (LM) with models incorporating text knowledge graph embeddings (KG). LR in the representation column denotes the linear regression learner and SGD denotes the Stochastic Gradient Descent learner.

Representation	Accuracy	F1 - score	Precision	Recall
LR(LM)	0.6200	0.6481	0.6034	0.7000
LR(KG)	0.6750	0.6859	0.6635	0.7100
LR(LM + KG)	0.6200	0.6481	0.6034	0.7000
SGD(LSA + TransE + RotatE)	0.7200	**0.7348**	**0.6900**	**0.7900**
SOTA (literature) (Buda & Bolonyai, 2020)	**0.7500**	x	x	x

5.1.3 Task 3: PAN2020

For the *PAN2020* problem, the best performing model uses the combination of the LSA document representation and the TransE and RotatE document representations and SGD based linear model on the subsets of all of the representations learned. The deeper neural networks failed to exploit the intermediate representations to a greater extent due to the lack of data examples. However, the problem benefited an increase in performance with the introduction of KG-backed representations, gaining 5.5% absolute improvement over the LM-only representation. The low amount of data available for training made the neural representations fail behind the subset of the linearly stacked ones. Such learning circumstances provide an opportunity for further exploration in the potential of methods for feature selection before including all features in the intermediate features. The SOTA model for this task (Buda & Bolonyai, 2020) based on an ensemble of linear classifiers built on top of n-gram and statistical features. The evaluation of the task with respect to the models is shown in Table 5.3.

5.1.4 Task 4: COVID-19

The text-based representation of the model outperformed the derived KG representation in terms of all of the metrics. However, the combined representation of the text and knowledge present, significantly improved the score, with the biggest gain from the joint-intermediate representations. The best-performing representation for this task was the one that was learned on the concatenated representation via SNN with 1024 nodes. This data set did not contain metadata information, so we ommitted the KG-ENITTY evaluation. The evaluation of the task with respect to the models is shown in Table 5.4. The proposed method of stacking ensembles of representations outscored all other representations for all of the problems. The gain in recall and precision is evident for every problem, since the introduction of conceptual knowledge informs the textual representations about the concepts and the context. The SOTA (Glazkova et al., 2020) model for this task uses an ensemble of multiple transformer models - BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and COVID-Twitter-BERT (Müller et al., 2020). The best-performing models were the ones that utilized the textual representations and the factual knowledge of concepts appearing in the data.

Table 5.4: Comparison of representations on the *COVID-19* data set without background knowledge (LM) with models incorporating text knowledge graph embeddings (KG). LR in the representation column denotes the linear regression learner and SNN denotes the Shallow Neural Network learner.

Representation	Accuracy	F1 - score	Precision	Recall
LR(LM)	0.9285	0.9320	0.9275	0.9366
LR(KG)	0.8379	0.8422	0.8582	0.8268
LR(LM+KG)	0.9369	0.9401	0.9347	0.9455
SNN(LM+KG)	**0.9570**	0.9569	**0.9533**	**0.9652**
SOTA (literature) (Glazkova et al., 2020)	x	**0.9869**	x	x

5.2 Qualitative Results

In the following section, we further explore the constructed multi-representation space. In Subsection 5.2.1, we are interested in whether it is possible to pinpoint which parts of the space were the most relevant for a given problem. In Subsection 5.2.2 we analyze how representative the concept matching is. In Subsection 5.2.3, we analyze whether predictions can be explained with the state-of-the-art explanation methods.

5.2.1 Relevant feature subspaces

We next present a procedure and the results for identifying the key feature subspaces, relevant for a given classification task. We extract such features via the use of *supervised feature ranking*, i.e. the process of prioritizing individual features with respect to a given target space. In this thesis, we considered mutual information-based ranking (Kraskov et al., 2011), as the considered spaces were very high-dimensional (in both dimensions). As individual features are mostly latent, and as such non-interpretable, we are interested in what proportion the top k features correspond to a given subspace (e.g., the proportion of BoW features). In this way, we assessed the relevance of a given feature subspace amongst the top features. For the purpose of investigating such subspace counts across different data sets, we present the radial plot-based visualization, shown in Figure 5.1. The radial plot represents the global top ranked feature subspaces. It can be observed that very different types of features correspond to different data sets. For example, the LSA- and statistics-based features were the most relevant for the AAAI data set, however irrelevant for the others. On the other hand, where the knowledge graph-based type of features was relevant, we can observe that multiple different KG-based representations are present. A possible explanation for such behavior is that, as shown in Table 3.1, methods are to some extent complementary with respect to their expressive power, and could hence capture similar patterns. Individual data sets are inspected in Figure 5.2. For different data sets, different subspaces were the most relevant. For example, for the *FakeNewsNet*, the DistMult and simplE-based representations of given entities were the most frequently observed types of features in top 200 features. This parameter was selected with the aim to capture only the top-ranked features – out of thousands of features, we hypothesize that amongst the top 200 key subspaces are represented. The simplE-based features were also the most relevant for the *LIAR-PANTS* data set. However, for the *AAAI-COVID19* data set, the statistical and LSA-based features were the most relevant. A similar situation can be observed for the *PAN2020* data set, where statistical features were the most relevant. The observed differences in ranks demonstrate the utility of multiple representations and their different

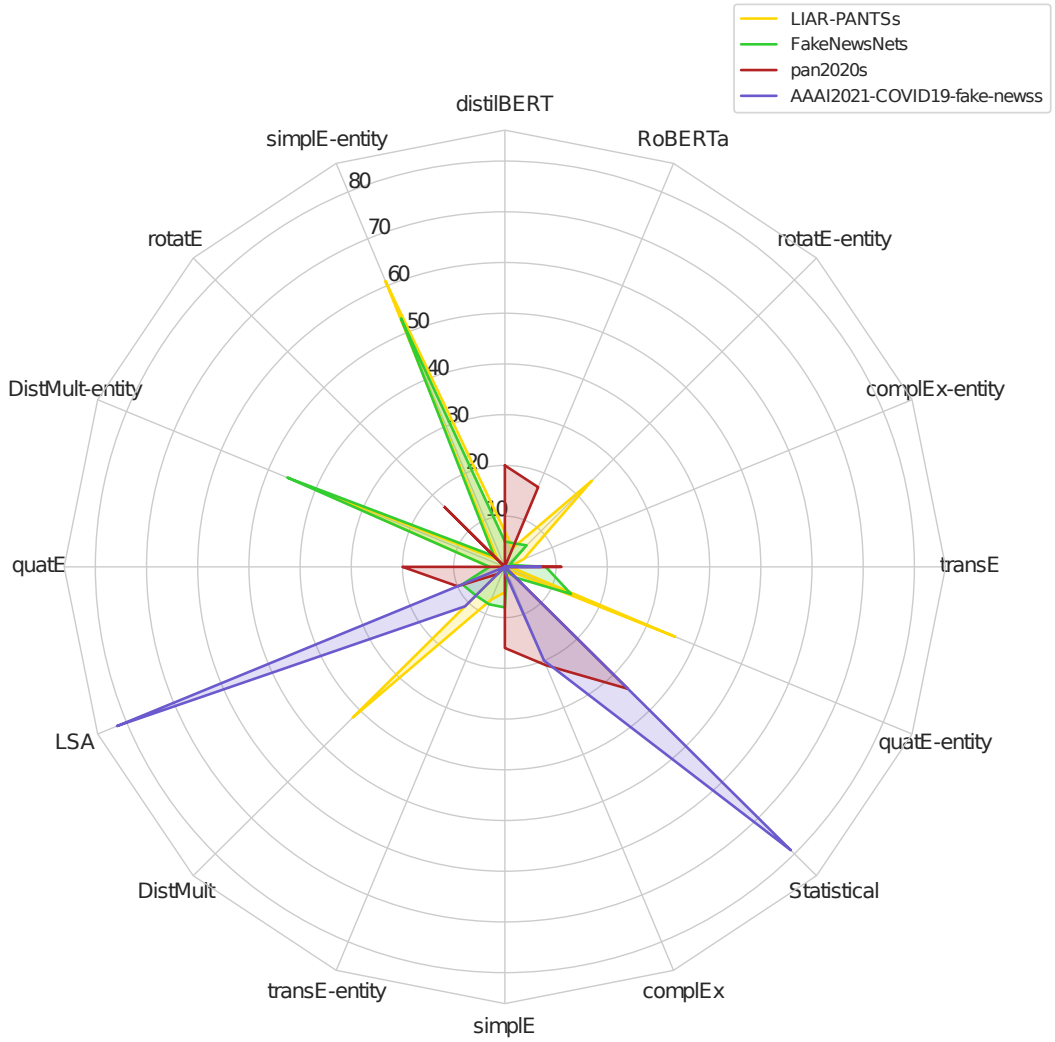


Figure 5.1: Overview of the most relevant feature subspaces for individual data sets.

relevance for individual classification tasks. By understanding the dominating features, one can detect general properties of individual data sets; e.g., high scores of statistical features indicate punctuation-level features could have played a prominent role in the classification. On the contrary, the dominance of entity embeddings indicates that semantic features are of higher relevance. Note that to our knowledge, this study is one of the first to propose the radial plot-based ranking counts as a method for global exploration of the relevance of individual feature subspaces.

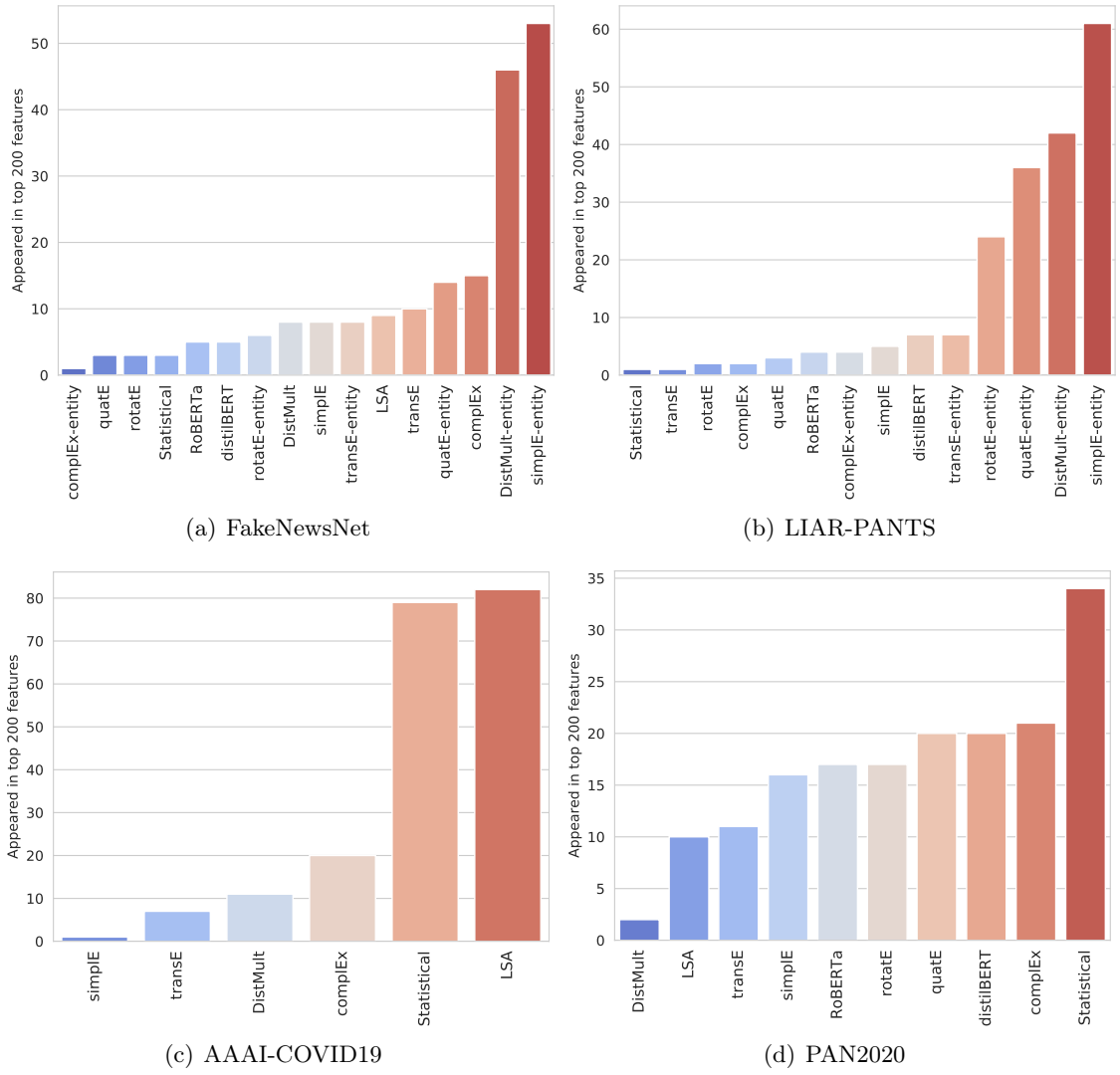
5.2.2 Exploratory data analysis study on the knowledge graph features from documents

In this section, we analyze how representative the concept matching is. As described in Subsection 3.2 for each document, we first generate the n-grams and extract those present in the KG. For each data set we present the top 10 most frequent concepts that were extracted. First we analyze the induced concepts for all four data sets, followed by the concepts derived from the document metadata for the *LIAR* and *FakeNewsNet* data set. The retrieved concepts are shown in Figure 5.3.

The data sets that focus on fake news in the political spectrum (*LIAR* and *FakeNewsNet*) appear to be described by concepts such as *government* and *governmental institutions*,

as well political topics revolving around *budget* and *healthcare*. In the case of the metadata representation *Donald Trump* and *Barack Obama* appear as most common. From the general metadata the political affiliation *democrat* comes out on top, followed by political topics such as *economy, taxes, elections* and *education*. Concepts related to the *coronavirus* such as *death, confirmed and reported cases, patients, pandemic, vaccine, hospital* appeared as the most representative in the *COVID-19* data set. Twitter posts are of limited length and of very versatile nature, making the most common concept in the *PAN2020* data set *URLs* to other sources. Following this, numbers and verbs describe the state of the author such as *need, give, could, and like*.

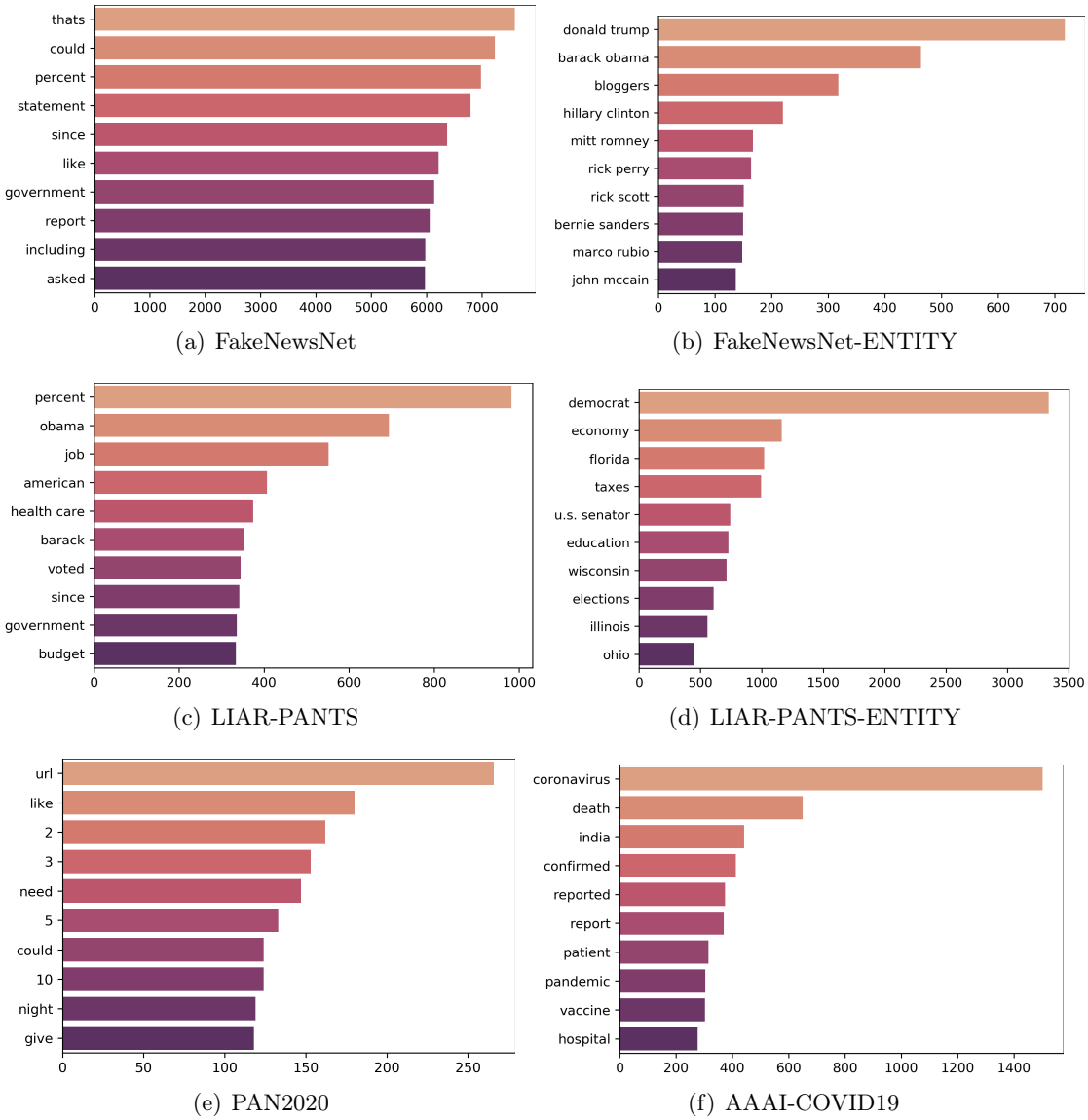
We finally discuss the different concepts that were identified as the most present across the data sets. Even though in data sets like FakeNewsNet and LIAR-PANTS, the most common concepts include well-defined entities such as e.g., 'job', the PAN2020 mapping indicates that this is not necessarily always the case. Given that only for this data set most



(a) FakeNewsNet (b) LIAR-PANTS

(c) AAAI-COVID19 (d) PAN2020

Figure 5.2: Inspection of ranked subspaces for individual data sets. Note that not all feature types are present amongst the top 200 features according to the feature ranking, indicating that for data sets like AAAI-COVID19, e.g., mostly LSA and statistical features are sufficient.



(a) FakeNewsNet (b) FakeNewsNet-ENTITY

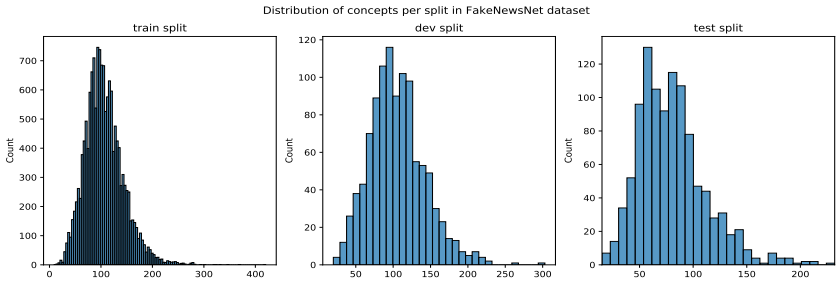
(c) LIAR-PANTS (d) LIAR-PANTS-ENTITY

(e) PAN2020 (f) AAAI-COVID19

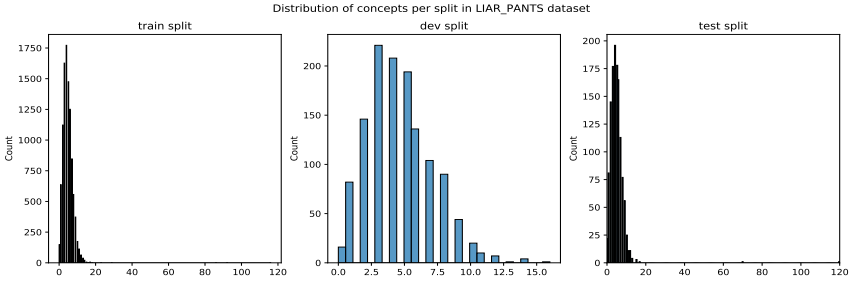
Figure 5.3: Most common concepts from the WikiData5m KG per article (training data) of the data sets. For the *FakeNewsNet* and *LIAR* data sets, we additionally report the most popular present concepts from the metadata. The x-axis reports the number of occurrences, while the y-axis reports the given concept.

frequent concepts also include, e.g., numbers, we can link this observation to the type of the data – noisy, short tweets. Having observed no significant performance decreases in this case, we conducted no additional denoising ablations (such as more intensive concept selection), even though such endeavor could be favourable in general.

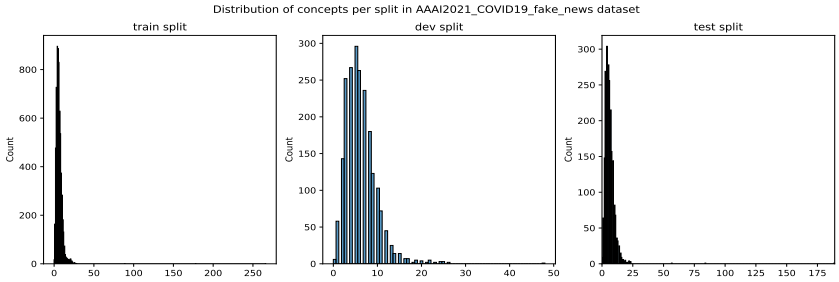
Next we analyze how much coverage of concepts per data set has the method acquired. We present the distribution of induced knowledge graph concepts per document for every data set in Figure 5.4. The number of found concepts is comparable across data sets. The chosen data sets have more than 98% of their instances covered by additional information, from one or more concepts. For the *LIAR* data set we fail to retrieve concepts only for 1.45% of the instances, for *COVID-19* only for 0.03% instances. In the case of *PAN2020* and *LIAR* data sets we succeed to provide one or more concepts for all examples.



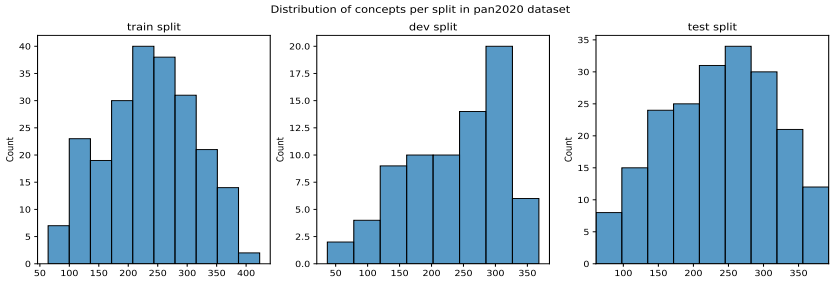
(a) FakeNewsNet



(b) LIAR-PANTS



(c) AAAI-COVID19



(d) PAN2020

Figure 5.4: Distribution of concepts extracted from the WikiData5m KG per article in the data sets.

5.2.3 Evaluation of word features in the data

To better understand data sets and obtained models, we inspected words in the *COVID-19 Fake News detection* set as features of the prediction model. We were interested in words that appeared in examples with different contexts which belonged to the same class. To find such words, we evaluated them with the TF-IDF measure, calculated the variance of these features separately for each class and extracted those with the highest variance in

their class.

We mapped the extracted words to WordNet (Fellbaum, 2010) and generalized them using Reasoning with Explanations (ReEx) (Perdih et al., 2021) to discover their hypernyms, which can serve as human understandable explanations.

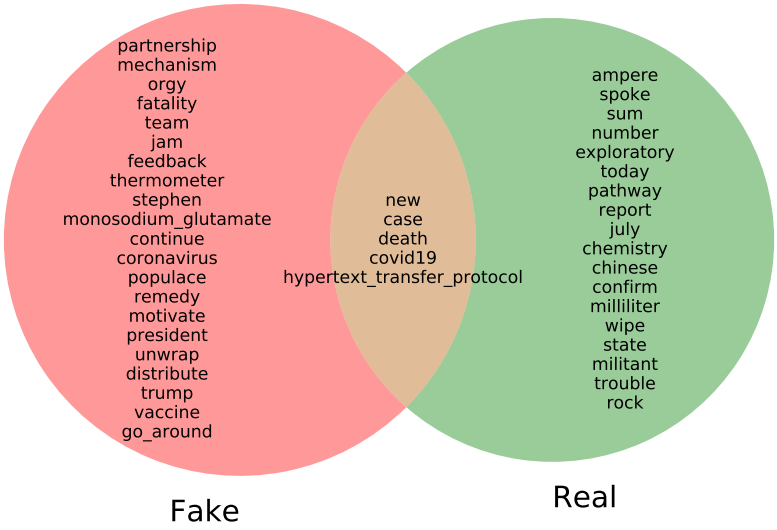


Figure 5.5: Words with the highest variance in their class. This is the first step towards providing understandable explanations of what affects the classification.

If examined separately, most words found based on variance offer very little as explanations. A couple of words stand out, however; since this is a COVID news data set, it is not surprising that words such as "new", "covid19", "death" and "case" are present across different news examples in both classes. Because COVID-19-related news and tweets from different people often contain contradictory information and statements, there must be fake news about vaccines and some substances among them, which could explain their inclusion among words appearing in examples belonging to the "fake" class. Words found in examples belonging to the "real" class seem to be more scientific and concerning measurements, for example, "ampere", "number", "milliliter". Figure 5.5 shows words with the highest variance in their respective class, while Figure 5.6 shows found hypernyms of words with the highest variance for each of the classes.

After generalizing words found with variance we can examine what those words have in common. "Causal agent" is a result of the generalization of words in both fake and real classes, which implies that news of both classes try to connect causes to certain events. These explanations also reveal that different measures, attributes and reports can be found in examples belonging to the "real" class.

5.2.4 Performance of individual feature spaces

We report the performances of individual representations presented as a part of this thesis next.

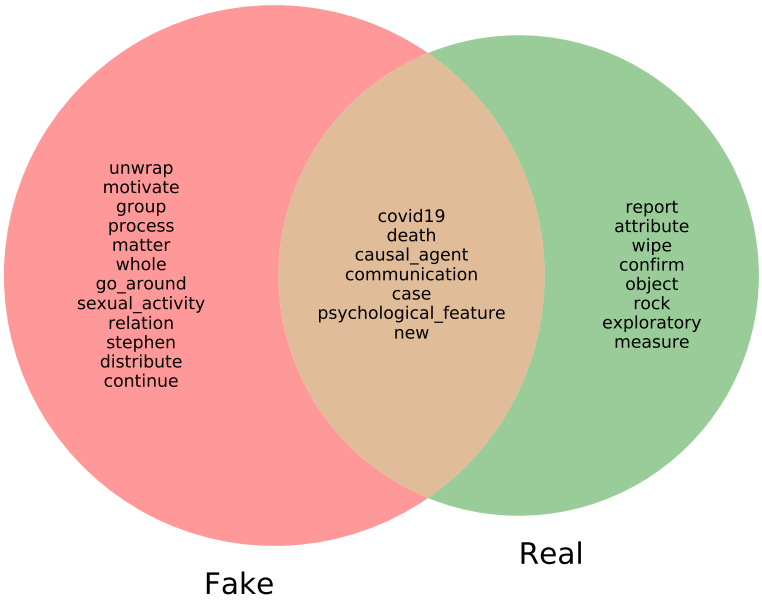


Figure 5.6: We used ReEx with Wordnet to generalize words with the highest variance in their class, and produce understandable explanations.

5.2.4.1 Evaluation of all subsets of spaces

In this section we explore how combining various spaces affect the performance. Due to the high cardinality of the document and knowledge-graph embedding we sample 10% with respect to the distribution of lables as in the original distribution. The only exception is the PAN2020 data set where we use the whole data set, due to the small number of examples. For every problem we evaluate all the possible combinations consisting of KG representations and LM, in all-in-all 11 representations making evaluated in total $2^{11}-1 = 2047$ combinations of features, on which we learn LogisticRegression classifier with various values of regularization $C \in \{1, 0.1, 0.01, 0.001\}$. For every problem we showcase the best 10 and the worst 10 combinations of features, evaluated at four different score techniques.

5.2.4.2 LIAR

The representations that captured only statistical and lexical features show low importance to the task when combined, resulting in an F1-score of 11.68%. The additional combination of lexical and contextual spaces provided improvement to the scores. The most significant gain on performance concerning the F1-score came with the combination of the QuatE and the simplE knowledge graph features with the dBERT model, improving the score by 11.42%. Multiple representations landed among the highest F1-score of 26.53%, the most interesting one is that the combination of DistilBERT and XLM model with statistical features and rotatE knowledge graph embedding yielded top performance. The dependence of the number of features and the F1-scores is represented in Figure 5.7. The worst-performing combinations are listed in Table 5.5, while the best-performing combinations

are listed in Table 5.6.

Table 5.5: LIAR worst 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
LSA_stat	522	0.116782	0.141732	0.117917	0.121464
rotate_roBERTa_stat_XLM	2058	0.127043	0.149606	0.127742	0.129400
rotate_LSA_roBERTa_stat_XLM	2570	0.127043	0.149606	0.127742	0.129400
transe_rotate_roBERTa_stat_XLM	2570	0.127043	0.149606	0.127742	0.129400
transe_rotate_LSA_roBERTa_stat_XLM	3082	0.127043	0.149606	0.127742	0.129400
transe_rotate_quate_distmult_simple_LSA	3072	0.131043	0.149606	0.137023	0.130886
rotate_quate_distmult_simple_LSA	2560	0.131043	0.149606	0.137023	0.130886
complex_rotate_quate_LSA_roBERTa_XLM	3584	0.134385	0.141732	0.139119	0.134308
LSA	512	0.137799	0.165354	0.138862	0.142240
complex_transe_rotate_quate_distmult_simple_LSA	3584	**0.137810**	**0.157480**	**0.143607**	**0.137337**

Table 5.6: LIAR best 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
transe_rotate_DistilBERT_LSA_XLM	3072	0.260089	0.275591	0.260826	0.261883
quate_simple_DistilBERT	1792	0.260485	0.275591	0.277576	0.257641
transe_quate_simple_DistilBERT	2304	0.260485	0.275591	0.277576	0.257641
rotate_DistilBERT_stat_XLM	2058	0.262555	0.275591	0.266784	0.262160
rotate_DistilBERT_LSA_stat_XLM	2570	0.262555	0.275591	0.266784	0.262160
transe_rotate_DistilBERT_LSA_stat_XLM	3082	0.262555	0.275591	0.266784	0.262160
transe_rotate_DistilBERT_stat_XLM	2570	0.262555	0.275591	0.266784	0.262160
complex_transe_quate_distmult_simple_DistilBERT_LSA_roBERTa	4608	0.265255	0.283465	0.269992	0.263042
complex_quate_distmult_simple_DistilBERT_roBERTa	3584	0.265255	0.283465	0.269992	0.263042
complex_transe_quate_distmult_simple_DistilBERT_roBERTa	4096	**0.265255**	**0.283465**	**0.269992**	**0.263042**

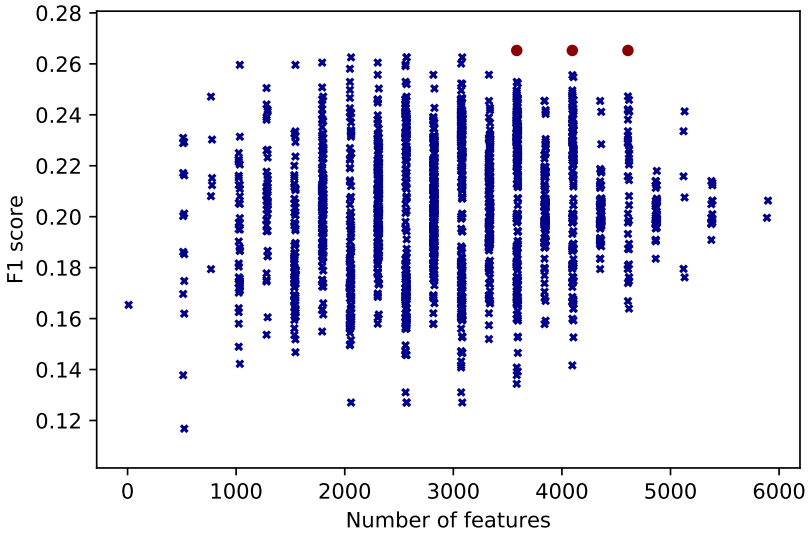


Figure 5.7: The interaction of dimensions and the F1-score for the LIAR problem. The red dots represent the highest scoring models.

5.2.4.3 FakeNewsNet

Knowledge graph and their combinations generated too general spaces that scored lowest on the data set. The lowest scoring representation is the one based only on the TransE KG embedding method. Notable improvement was seen with introduction of the contextual

Table 5.7: FakeNewsNet worst 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
transe	512	0.524066	0.528302	0.582348	0.572545
rotate_stat_XLM	1290	0.545714	0.547170	0.557471	0.559524
rotate_LSA_stat_XLM	1802	0.546524	0.547170	0.561957	0.563616
transe_rotate_LSA_stat_XLM	2314	0.546524	0.547170	0.561957	0.563616
transe_rotate_stat_XLM	1802	0.553384	0.556604	0.560606	0.563244
transe_rotate_quate_LSA_stat_XLM	2826	0.556248	0.556604	0.573953	0.575521
transe_rotate_quate_distmult_stat_XLM	2826	0.556564	0.556604	0.584428	0.583705
rotate_XLM	1280	0.563552	0.566038	0.572143	0.575149
transe_distmult_XLM	1792	0.563552	0.566038	0.572143	0.575149
rotate_quate_distmult_stat_XLM	2314	**0.566038**	**0.566038**	**0.591518**	**0.591518**

Table 5.8: FakeNewsNet best 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
complex_LSA_roBERTa_XLM	2560	0.753312	0.754717	0.761429	0.772321
transe_rotate_quate_distmult_roBERTa_XLM	3584	0.753312	0.754717	0.761429	0.772321
transe_rotate_simple	1536	0.754630	0.754717	0.780425	0.784598
complex_rotate_quate	1536	0.754717	0.754717	0.788690	0.788690
complex_transe_rotate_simple_LSA	2560	0.754717	0.754717	0.788690	0.788690
complex_rotate_quate_simple_LSA	2560	0.754717	0.754717	0.788690	0.788690
complex_rotate_stat	1034	0.773262	0.773585	0.792391	0.800223
complex_transe_simple_LSA	2048	0.773585	0.773585	0.808408	0.808408
complex_simple_LSA	1536	0.773585	0.773585	0.808408	0.808408
complex_transe_rotate_stat	1546	**0.782535**	**0.783019**	**0.798594**	**0.808036**

representation. The best performing model for this problem was the one that combined features from knowledge graphs that preserve various relations(the ComplEx, TransE, and RotatE embeddings) and the simple stylometric representation. The dependence of the number of features and the F1-scores is represented in Figure 5.8. The worst-performing combinations are listed in Table 5.7, while the best-performing combinations are listed in Table 5.8.

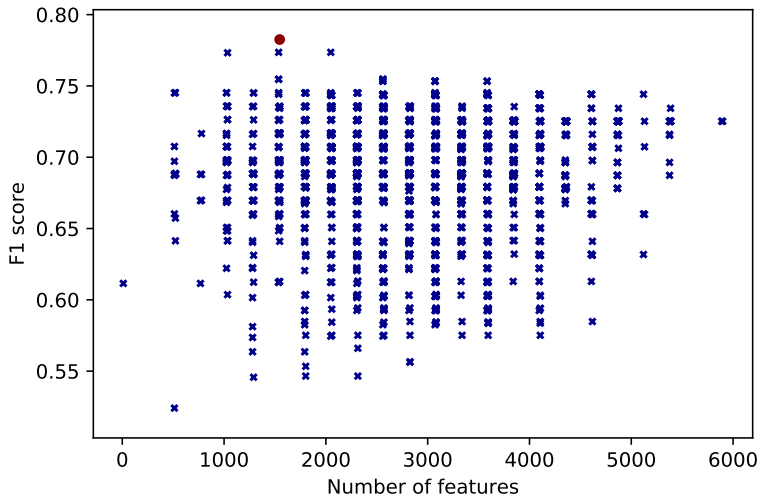


Figure 5.8: The interaction of dimensions and the F1-score for the FakeNewsNet problem. The red dots represent the highest scoring models.

5.2.4.4 PAN2020

For the PAN2020 problem, the combination of the knowledge graph representations with the contextual-based language representations as XLM ranked the lowest, with a F1-score of 57.45%. The problem benefited the most from the LSA representation, the additional enrichment of this space with knowledge graph features improved the score by 14.02%. The best-performing model based on ComplEx and QuatE KG embeddings and LSA and statistical language features, with a dimension of 1546. The worst-performing combinations are listed in Table 5.9, while the best-performing combinations are listed in Table 5.10. The dependence of the number of features and the F1-scores is represented in Figure 5.9.

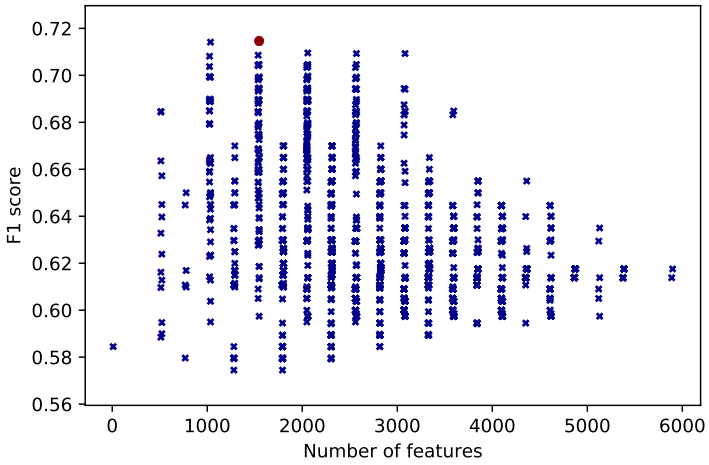


Figure 5.9: The interaction of dimensions and the F1-score for the PAN2020 problem. The red dots represent the highest scoring models.

Table 5.9: PAN2020 worst 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
complex_transe_XLM	1792	0.574479	0.575	0.575369	0.575
complex_XLM	1280	0.574479	0.575	0.575369	0.575
quate_LSA_XLM	1792	0.579327	0.580	0.580515	0.580
quate_distmult_XLM	1792	0.579327	0.580	0.580515	0.580
transe_quate_distmult_XLM	2304	0.579327	0.580	0.580515	0.580
transe_quate_LSA_XLM	2304	0.579327	0.580	0.580515	0.580
transe_LSA_XLM	1792	0.579327	0.580	0.580515	0.580
complex_transe_LSA_XLM	2304	0.579327	0.580	0.580515	0.580
complex_LSA_XLM	1792	0.579327	0.580	0.580515	0.580
LSA_XLM	1280	**0.579327**	**0.580**	**0.580515**	**0.580**

Table 5.10: PAN2020 best 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
complex_transe_quate_distmult_LSA_stat	2570	0.704638	0.705	0.706009	0.705
complex_quate_distmult_LSA_stat	2058	0.704638	0.705	0.706009	0.705
distmult_LSA	1024	0.708132	0.710	0.715517	0.710
transe_distmult_LSA	1536	0.708572	0.710	0.714198	0.710
complex_transe_quate_distmult_simple_LSA_stat	3082	0.709273	0.710	0.712121	0.710
complex_quate_distmult_simple_LSA_stat	2570	0.709273	0.710	0.712121	0.710
complex_transe_quate_LSA_stat	2058	0.709535	0.710	0.711353	0.710
transe_quate_LSA_stat	1546	0.714135	0.715	0.717633	0.715
quate_LSA_stat	1034	0.714135	0.715	0.717633	0.715
complex_quate_LSA_stat	1546	**0.714650**	**0.715**	**0.716059**	**0.715**

5.2.4.5 COVID-19

Knowledge-graph-only-based representation yielded too general spaces, resulting in the lowest-performing spaces for the COVID-19 task. Notable improvement for the data set was achieved by the addition of language models to the knowledge graph representations. The worst-performing combinations are listed in Table 5.11, while the best-performing combinations are listed in Table 5.12. The dependence of the number of features and the F1-scores is represented in Figure 5.10.

Table 5.11: COVID-19 worst 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
complex_transe_distmult	1536	0.695936	0.696262	0.695893	0.696254
complex_distmult	1024	0.695936	0.696262	0.695893	0.696254
complex_transe_rotate_quate_distmult	2560	0.705447	0.705607	0.705607	0.706057
transe_rotate_distmult	1536	0.709875	0.710280	0.709790	0.710084
complex_rotate_quate_distmult	2048	0.710179	0.710280	0.710517	0.710959
rotate_distmult	1024	0.724004	0.724299	0.723941	0.724352
complex	512	0.724293	0.724299	0.725488	0.725665
complex_quate_distmult	1536	0.728379	0.728972	0.728379	0.728379
complex_transe_quate_distmult	2048	0.728379	0.728972	0.728379	0.728379
transe_rotate_quate_distmult	2048	**0.728593**	**0.728972**	**0.728497**	**0.728817**

Table 5.12: COVID-19 best 10 representation combinations.

combination	dimensions	f1_scores	accuracy_score	precision_score	recall_score
transe_rotate_quate_simple_DistilBERT_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
transe_rotate_distmult_simple_DistilBERT_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
transe_quate_distmult_simple_DistilBERT_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
rotate_quate_distmult_simple_DistilBERT_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
rotate_quate_distmult_DistilBERT_LSA_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
rotate_distmult_simple_DistilBERT_LSA_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
transe_rotate_quate_distmult_simple_DistilBERT_LSA_roBERTa	4608	0.910886	0.911215	0.911770	0.910364
complex_transe_rotate_quate_distmult_DistilBERT_roBERTa	4096	0.910886	0.911215	0.911770	0.910364
complex_distmult_simple_DistilBERT_LSA_roBERTa	3584	0.910886	0.911215	0.911770	0.910364
LSA	512	**0.911058**	**0.911215**	**0.910916**	**0.911239**

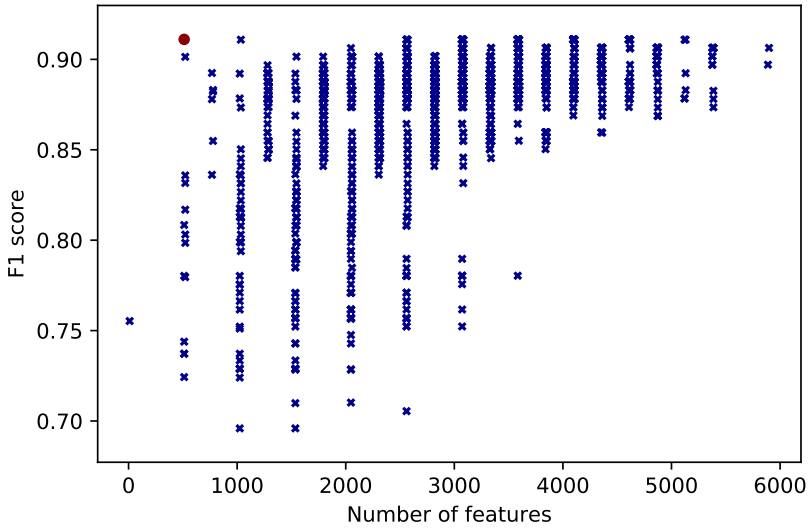


Figure 5.10: The interaction of dimensions and the F1-score for the COVID-19 problem. The red dots represent the highest scoring models.

5.2.5 Conclusion on qualitative evaluation

Based on the ablation studies from Sections 5.2.1, 5.2.2, 5.2.3, 5.2.4, targeting the performance of different feature space combinations, there are two main takeaways:

1. Knowledge-graph-based representations on their own are too general for fake news detection tasks, where the main type of input are short texts. However, combining knowledge-graphs with additional statistical and contextual information about such texts has shown to improve the performance. The representations that are capable of capturing different types of relation properties (e.g., symmetry, asymmetry, inversion etc.) in general perform better than the others.

2. We observed no general rule determining the optimal representation combination. Current results, however, indicate that transfer learning based on different representation types is a potentially interesting research direction. Furthermore, similarity between the spaces could be further studied at the task level.

Chapter 6

Conclusions and Further Work

We compared different representations methods for text, graphs and concepts, and proposed a novel method for merging them into a more efficient representation for the detection of fake news. We analysed statistical features, matrix factorization embedding LSA, and neural sentence representations sentence-bert, XLM, dBERT, and RoBERTa. We proposed a concept enrichment method for document representations based on data from the WikiData5m knowledge graph. The proposed representations significantly improve the model expressiveness and improve classification performance in all addressed tasks.

The fake news problem space captured in the aforementioned data sets showed that no single representation or an ensemble of representation works consistently for all problems – different representation ensembles improve performance for different problems. For instance the author profiling - PAN2020 problem gained performance increase from only a subset of representations i.e. the *TransE* and *SimplE* KG derived concepts. As for the FakeNewsNet, the best-performing model was a heterogeneous ensemble of all of the constructed representations and the metadata representations.

The evaluation of the proposed method also showed that the KG-only representations were good enough in the case of *PAN2020, LIAR and COVID-19*, where they outperformed the text-only based representations. This represents a potential of researching models based both on contextual and factual knowledge while learning the language model. Z. Wang et al. (2014) reported that such approaches can introduce significant improvement; with the increase of the newer methods and mechanisms popular in NLP today we believe this is a promising research venue.

The solutions to some problems benefit from some properties while others benefit from others, in order to explore the possibility one can perform a search through the space of combinations of the available KG models. However, exhaustive search can introduce significant increase in the memory and time complexity of learning models. One way to cope with this problem is to apply some regularization to the learner model which would learn on the whole space. The goal of this would be to omit the insignificant combinations of features to affect the predictions of the model. Another approach would be to perform feature selection and afterwards learn only on the representations that appear in the top k representative features.

The drawbacks of the proposed method include the memory consumption and the growth of the computational complexity with the introduction of high dimensional spaces. In order to cope with the curse of dimensionality issue we propose exploring some dimensionality-reduction approaches such as UMAP (McInnes et al., 2018) that map the original space to a low-dimensional manifold (Angelov, 2020). Another problem of the method is that it is not capable of choosing the right approach for concept extraction from a given text. In the current work for the fuzzy matching, we consider every n-gram present in the document

and the KG as a possible candidate. For the further work we consider exploring how the performance will benefit if we apply some entity extraction tools and extract only entities of particular interest. Furthermore, a potential drawback of the proposed method is relatively restrictive entity-to-document mapping. By adopting some form of fuzzy matching, we believe we could improve the mapping quality and with it the resulting representations.

For further work we also propose exploring attention-based mechanisms to derive explanations for the feature significance of a classification of an instance. Additionally we would like to explore more advanced concept aggregation weighting schemes, such as the *AGG-TF* where the frequency of appearance of a given concepts through the document will be taken into account or the *AGG-TF-IDF* where a term weighting would be leveraged by the TF-IDF weight of that term. The intensive amount of research focused on the Graph Neural Networks represents another potential field for exploring our method. The combination of different KG embeddings (such as TuckER (Balažević et al., 2019) or Multi-relational Poincaré Graph embeddings (Balaževič et al., 2019)) would capture more different patterns and therefore improve the knowledge-aware representations and aid the heterogeneous representations.

In this thesis we omitted the social-context (interaction of users, time-stamp, network information, etc), which is highly relevant to the fake news problem. In future work, we plan to experiment on the network of fake news spreading. We will consider using the heterogeneous knowledge aware representations as node information and utilizing graph neural networks to learn the final embeddings.

We also hypothesize that the knowledge representations can be utilized in a multi-task or transfer learning scenario, since they are founded on factual truth that cannot be interpreted ambiguously. We propose learning to solve multiple fake news tasks simultaneously in a multi-task setting. Since fake news detection data sets can originate from different domains, we believe that by leveraging common knowledge in the learning, we will be able to contribute to a domain agnostic fake news detection model.

Transfer learning of knowledge learned specific models on a language or task level is currently a hot topic within the NLP research community. In this thesis we only consider solving the tasks in English, however for further work we want to consider learning both the document and knowledge aware representations from multilingual sources. Next step in such an experimental scenario would be evaluating the proposed representations in a cross-lingual setting.

The content of this thesis was published in the journal paper (Koloski et al., 2022). The code is freely accessible at https://github.com/bkolosk1/KBNR.

Chapter 7

Related Publications

In this chapter, we list the publications related to the method in this work. The chapter consists of two parts - the first focuses on the *Fake news detection* methods that led to the development of this work and the second focuses on works that implemented and used the proposed method.

7.1 Fake news detection

In this section, we describe two papers leading to this work and the final version of this work published in a journal.

7.1.1 Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization

This contribution explored the *detection of fake news spreaders in multilingual settings* backed by features based on latent semantic analysis (LSA). In this work, we tackled the problem of identifying twitter authors as potential spreaders of fake news. For each author we were given *100* tweets and a corresponding label - denoting the author as *spreader* or *non-spreader*. The problem was proposed in two languages: English and Spanish, with same class distribution. We treated this problem as document classification. To convert the problem we first concatenated all the tweets of a given tweet into a single document. For document representation, we focused on TF-IDF weighted word and character n-grams that where later transformed into a Latent Semantic Analysis space via singular-value-decomposition **SVD**.

CEUR-WS.org/Vol-2696/paper_145.pdf

Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization
Notebook for PAN at CLEF 2020

Boško Koloski[1,2], Senja Pollak[1], and Blaž Škrlj[1]

[1]Jožef Stefan Institute, Ljubljana
[2]Faculty of Information Science - University of Ljubljana , Slovenia
blaz.skrlj@ijs.si

Abstract Fake news is an emerging problem in online news and social media. Efficient detection of fake news spreaders and spurious accounts across multiple languages is becoming an interesting research problem, and is the key focus of this paper. Our proposed solution to PAN 2020 fake news spreaders challenge models the accounts responsible for spreading the fake news by accounting for different types of textual features, decomposed via sparse matrix factorization, to obtain easy-to-learn-from, compact representations, including the information from multiple languages. The key contribution of this work is the exploration of how powerful and scalable matrix factorization-based classification can be in a multilingual setting, where the learner is presented with the data from multiple languages simultaneously. Finally, we explore the joint latent space, where patterns from individual languages are maintained. The proposed approach scored second on the 2020 PAN shared task for identification of fake news spreaders.

1 Introduction

The notion of fake news refers to distortions of news with the intention to affect the political landscape and to create confusion and divisions in society. Even if the phenomenon of fake news is not new, the scale and impact of fake news has never been so important than today, which can be attributed to the digital transformation of the news industry, and especially to the rise of social media as a news distribution channel. [6]

One of the crucial problems is the recognition of *fake news spreaders*. For example, Twitter bots (fake accounts) are capable of generating fake information and propagating it through their follower networks, which can impact real-life entities such as stock markets and possibly even elections [4]. Automatic detection of such spreaders is thus becoming one of the key approaches to minimize the manual annotation costs employed by the social media owners. This work fits under the framework of the PAN author profiling tasks [21,19], and describes our approach submitted to the PAN 2020 shared task on Profiling Fake News Spreaders on Twitter [22].

On top of the derived representation we used Stochastic Gradient Based learner with both Logistic Regression and SVM kernels. We experimented with both monolingual representations for each language and a joint multilingual representation. The multilingual representation outperformed the monolingual representation on both - the internal evaluation and the test evaluation. The proposed solution achieved F1-score of 0.7550, ranking second out of 33 entries at the shared task[1].

[1]https://pan.webis.de/clef20/pan20-web/author-profiling.html#Results

Koloski, B., Pollak, S., & Škrlj, B. (2020). Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org

7.1.2 Identification of COVID-19-related Fake News via Neural Stacking

The following publication was our foundational work on the development of *heterogeneous document representations*. We tackled the task of identifying a single social media post as possible COVID19 related fake news. Initially, we relayed simple stylometric features based on word and character statistics such as minima, maxima and average frequency of occurrence. Next, we focused on inclusion of more sophisticated features like LSA and contextually rich features like contextual BERT representations.

For contextual representations we utilized the sentence-transformers variants of distilBERT Sanh et al. (2019), RoBERTa Liu et al. (2019) and XLM Conneau and Lample (2019). Similarly as in 7.1.1, we used StochasticGradient-based learner with both Logistic Regression and SVM kernels. In addition, we learned end-to-end representation learners and classifiers like tax2vec Škrlj et al., 2021 and the hugging-face variant of the distilBERT Sanh et al., 2019 model.

Identification of COVID-19 Related Fake News via Neural Stacking

Boshko Koloski[1,2(0)], Timen Stepišnik-Perdih[3], Senja Pollak[1], and Blaž Škrlj[1,2]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[boshko.koloski,blaz.skrlj]@ijs.si
[2] Jožef Stefan Int. Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia

Abstract. Identification of Fake News plays a prominent role in the ongoing pandemic, impacting multiple aspects of day-to-day life. In this work we present a solution to the shared task titled *COVID19 Fake News Detection in English*, scoring the 50th place amongst 168 submissions. The solution was within 1.5% of the best performing solution. The proposed solution employs a heterogeneous representation ensemble, adapted for the classification task via an additional neural classification head comprised of multiple hidden layers. The paper consists of detailed ablation studies further displaying the proposed method's behavior and possible implications. The solution is freely available. https://github.com/boshko/koloski/covid19-fake-news

Keywords: Fake-news detection · Stacking ensembles · Representation learning

1 Introduction

Fake news can have devastating impact on the society. In the times of a pandemic, each piece of information can have a significant role in the lives of everyone. The verification of the truthfulness of a given information as a fake or real is crucial, and can be to some extent learned [10]. Computers, in order to be able to solve this task, need the data represented in a numeric format in order to draw patterns and decisions. We propose a solution to this problem by employing various natural language processing and learning techniques.

The remainder of this work is structured as follows: Sect. 2 describes the prior work in the field of detection of fake-news. The provided data is described in Sect. 3 and Sect. 4 explains our proposed problem representation approaches while Sect. 5 introduces two different meta-models built on top of the basic representations listed in Sect. 4. The experiments and results achieved are listed in Sect. 6, finally the conclusion and the proposed future work are listed in Sect. 7.

© Springer Nature Switzerland AG 2021.
T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 177–188, 2021.
https://doi.org/10.1007/978-3-030-73696-5_17

We considered creating *heterogeneous representations via neural stacking* in two different paradigms. For the first one, we combined the outputs of the standalone models for each representations into a combined ensemble of classifiers. For the second, we considered stacking individual representations into a single one, and proceed to learn a deep neural joint *heterogeneous* representation. The **joint heterogeneous representations** offered the best performance, with test F1-score of 97.2% falling behind the top solution only by 1.5%.

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 177–188

7.1.3 Knowledge graph informed fake news classification via heterogeneous representation ensembles

In this final related representation we combine the aforementioned approaches and propose a new knowledge-based approach for representation. This work represents a follow-up of the two previous works with additional knowledge-derived representations. We utilize the Wikipedia5m knowledge-graph for leveraging out knowlede-graph representation. We used six different knowledge-graph entity embeddings to transform the concepts from the knowledge-graph into a numerically representative units. Next, for each KG concept present in the document we retrieve its embedding. Finally, we average the retrieved embeddings and obtain the entity aware document representation.

ARTICLE IN PRESS

Neurocomputing

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Knowledge graph informed fake news classification via heterogeneous representation ensembles

Boshko Koloski[a,b], Timen Stepišnik Perdih[b], Marko Robnik-Šikonja[c], Senja Pollak[a], Blaž Škrlj[a,b,*]

[a] Jožef Stefan Int. Postgraduate School, Slovenia
[b] Jožef Stefan Institute, 1000 Ljubljana, Slovenia
[c] University of Ljubljana, Faculty of Computer and Information Science, 1000 Ljubljana, Slovenia

ARTICLE INFO

ABSTRACT

1. Introduction

We proceed to learn the final *heterogeneous representations* via deep neural network where the inputs are the aforementioned representations. We benchmarked the proposed approach on five standard datasets for fake news detection. The method ranked on par with the state-of-the-art models while being less computationally heavy. In one fake news detection task the method achieved a state-of-the-art result.

> Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, *496*, 208–226. https://doi.org/https://doi.org/10.1016/j.neucom.2022.01.096

7.2 Application of the Method on Other Domains

In this section, we describe the impact of this method as it was applied to two different tasks and achieved competitive results with the best-performing methods.

7.2.1 E8-IJS@LT-EDI-ACL2022-BERT, AutoML and Knowledge-Graph Backed Detection of Depression

The subsequent related publication concerns the task of depression detection. In this work we were tasked to design a system to detect depression given a social media post. We followed the approach proposed in this work to derive the document representations. A novelty in this work is the way of constructing the heterogeneous representations. We first concatenate the different types of representation together into a single heterogeneous representation. In the final step, we considered the singular value decomposition on the stacked representations to obtain the **latent heterogeneous representations**. We learned a logistic regression as a classifier for this task. The latent stacked representations outscored the ones that were obtained via neural stacking on the internal experiments. The proposed approach ranked 8*th* out of 33 places in terms of F1-score, while it ranked 5*th* out of 33 in terms of recall. The main conclusion from this work is that for different problems, different ways of stacking should be explored.

E8-IJS@LT-EDI-ACL2022 - BERT, AutoML and Knowledge-graph backed Detection of Depression

Ilija Tavchioski* Jožef Stefan Institute Jamova 39, Ljubljana liljatavchioski@gmail.com

Boshko Koloski* Jožef Stefan Institute Jožef Stefan Institute IPS Ljubljana, Slovenia boshko.koloski@ijs.si

Blaž Škrlj* Jožef Stefan Institute Jamova 39, Ljubljana blaz.skrlj@ijs.si

Senja Pollak Jožef Stefan Institute Jamova 39, Ljubljana senja.pollak@ijs.si

Abstract

Depression is a mental illness that negatively affects a person's well-being and care, if left untreated, lead to serious consequences such as suicide. Therefore, it is important to recognize the signs of depression early. In the last decade, social media has become one of the most common places to express one's feelings. Hence, there is a possibility of text processing and applying machine learning techniques to detect possible signs of depression. In this paper, we present our approaches to solving the shared task titled *Detecting Signs of Depression from Social Media Text*. We explore three different approaches to solve the challenge; fine-tuning BERT model, leveraging AutoML for the construction of features and classifier selection and finally, we explore latent spaces derived from the combinations of textual and knowledge-based representations. We ranked 8th out of 33 teams in the competition, and our best solution, based on knowledge graph and textual embeddings, was only 0.1% behind the best model in terms of Macro F1, and only 1.9% behind in terms of Recall.

1 Introduction

Depression is a type of mental illness that affects a large part of our society and is one of the most complex challenges facing our humanity. Since depression is a disease that, if left untreated, can lead to serious consequences such as suicide over time, its early detection is crucial. Since people with depression typically do not open up in person very often, they often use social media as a way to express their thoughts and feelings (Steger and Kashdan, 2009). This trend increased rapidly with the COVID-19 pandemic due to restrictive measures that encouraged people to use social media as a means of expression. As the number of posts on social media has increased rapidly in recent years, there is a need to process them automatically to extract valuable information such as signs of depression. For this task, detecting signs of depression is a standard multi-class classification problem, where each post can be assigned to one of three classes ("non-depressive", "moderate", and "severe"). Increasing predictive accuracy may be critical for psychiatrists to detect the early signs of major depression to prevent further consequences. The remainder of this article is organized as follows; in Section 2, we discuss approaches to solve multi-class text classification problems and related work with data sets of social media posts on depression, in Section 3, we present the statistics of our given data, in Section 4, we explain the methods we used to solve the given problem, in Section 5, we present and analyze the results, and in Section 6, we provide the conclusion and present our plans for future work.

2 Related work

In related work, we can find various approaches to detecting depression from textual data including various social media.

One of the frequently used sources is **Twitter**. In one of the earlier studies of data from Twitter users who have attempted to take their life, Coppersmith et al. (2016) propose a logistic regression classifier using character *n*-gram character features. Leis et al. (2019) tackled the task of detecting depression in Spanish tweets and used occurrences of negative and positive words determined by Spanish Sentiment Lexicon (Pérez-Rosas et al., 2012) and

* Equal contribution.

> Tavchioski, I., Koloski, B., Škrlj, B., & Pollak, S. (2022). E8-IJS@ LT-EDI-ACL2022-BERT, AutoML and Knowledge-graph backed Detection of Depression. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 251–257

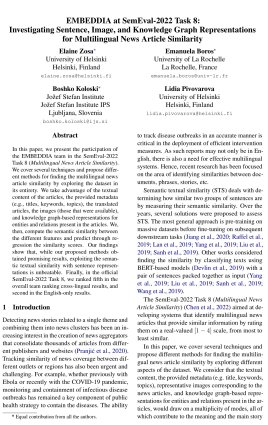
7.2.2 EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity

The final related publication represents the attempt to apply the proposed approach in the task of article similarity assessment. The given problem assesses the similarity of two news articles (not necessarily both being in the same language), via ranking from 1-5, with 1 being very similar and 5 being very different. In this paper, we considered using various modalities starting from the meta information of a given article such as keywords,

present images, contemporary document representations as BERT and knowledge-based representations.

To solve the issue of multilingual articles we first automatically translated the articles to English, and proceeded with the knowledge-graph embedding of the articles. We built standalone models based on the knowledge-backed representation and fused models with the other modalities. However, possibly due to automatic translation, our knowledge-graph ranking was fuzzy and thus lagged behind the top-performing models by a small margin.

Zosa, E., Boros, E., Koloski, B., & Pivovarova, L. (2022). EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity



EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity

Elaine Zosa*
University of Helsinki
Helsinki, Finland
elaine.zosa@helsinki.fi

Emanuela Boros*
University of La Rochelle
La Rochelle, France
emanuela.boros@univ-lr.fr

Boshko Koloski*
Jožef Stefan Institute
Jožef Stefan Institute IPS
Ljubljana, Slovenia
boshko.koloski@ijs.si

Lidia Pivovarova
University of Helsinki
Helsinki, Finland
lidia.pivovarova@helsinki.fi

Abstract

In this paper, we present the participation of the EMBEDDIA team in the SemEval-2022 Task 8 (*Multilingual News Article Similarity*). We cover several techniques and propose different methods for finding the multilingual news article similarity by exploring the dataset in its entirety. We take advantage of the textual content of the articles, the provided metadata (e.g., titles, keywords, topics), the translated articles, the images (those that were available), and knowledge graph-based representations for entities and relations present in the articles. We, then, compare the semantic similarity between the different features and predict through regression the similarity scores. Our findings show that, while our proposed methods obtained promising results, exploiting the semantic textual similarity with sentence representations is unbeatable. Finally, in the official SemEval-2022 Task 8, we ranked 48th in the overall team ranking cross-lingual results, and second in the English-only results.

1 Introduction

Detecting news stories related to a single theme and combining them into news clusters has been an increasing interest in the creation of news aggregators that consolidate thousands of articles from different publishers and websites (Pranjić et al., 2020). Tracking similarity of news coverage between different outlets or regions has also been urgent and challenging. For example, whether previously with Ebola or recently with the COVID-19 pandemic, monitoring and containment of infectious disease outbreaks has remained a key component of public health strategy to contain the diseases. The ability to track disease outbreaks in an accurate manner is critical in the deployment of efficient intervention measures. As such reports may not only be in English, there is also a need for effective multilingual systems. Hence, recent research has been focused on the area of identifying similarities between documents, phrases, stories, etc.

Semantic textual similarity (STS) deals with determining how similar two groups of sentences are by measuring their semantic similarity. Over the years, several solutions were proposed to assess STS. The most general approach is pre-training on massive datasets before fine-tuning on subsequent downstream tasks (Jiang et al., 2020; Raffel et al., 2019; Lan et al., 2019; Yang et al., 2019; Liu et al., 2019; Sanh et al., 2019). Other works considered finding the similarity by classifying texts using BERT-based models (Devlin et al., 2019) with a pair of sentences packed together as input (Yang et al., 2019; Liu et al., 2019; Sanh et al., 2019; Wang et al., 2019).

The SemEval-2022 Task 8 (*Multilingual News Article Similarity*) (Chen et al., 2022) aimed at developing systems that identify multilingual news articles that provide similar information by rating them on a real-valued $[1 - 4]$ scale, from most to least similar.

In this paper, we cover several techniques and propose different methods for finding the multilingual news article similarity by exploring different aspects of the dataset. We consider that the textual content, the provided metadata (e.g. title, keywords, topics), representative images corresponding to the news articles, and knowledge graph-based representations for entities and relations present in the articles, would draw on a multiplicity of modes, all of which contribute to the meaning and the main story

* Equal contribution from all the authors.

References

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modeling. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90. https://doi.org/10.18653/v1/W18-5513

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211–36.

Amirkhani, F. S. A. J. B. H. (2020). FNID: Fake news inference dataset. https://doi.org/10.21227/fbzd-sw81

Angelov, D. (2020). Top2vec: Distributed representations of topics.

Balažević, I., Allen, C., & Hospedales, T. M. (2019). Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*.

Balažević, I., Allen, C., & Hospedales, T. M. (2019). Multi-relational Poincaré Graph embeddings. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 4465–4475).

Balouchzahi, F., Shashirekha, H., & Sidorov, G. (2021). Mucic at checkthat! 2021: Fadofake news detection and domain identification using transformers ensembling. *Faggioli et al.[33]*.

Bidgoly, A., Amirkhani, H., & Sadeghi, F. (2020). Fake news detection on social media using a natural language inference approach.

Bordes, A., Usunier, N., Garcıéa-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013.* (pp. 2787–2795).

Buda, J., & Bolonyai, F. (2020). An Ensemble Model Using N-grams and Statistical Featuresto Identify Fake News Spreaders on Twitter—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Chandra, S., Mishra, P., Yannakoudakis, H., Nimishakavi, M., Saeidi, M., & Shutova, E. (2020). Graph-based modeling of online communities for fake news detection. *CoRR*, *abs/2008.06274*.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 7057–7067).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dun, Y., Tu, K., Chen, C., Hou, C., & Yuan, X. (2021). KAN: Knowledge-aware attention network for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), 81–89.

Fei, H., Ren, Y., Zhang, Y., Ji, D., & Liang, X. (2020). Enriching contextualized language model from knowledge graph for biomedical information extraction [bbaa110]. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa110

Fellbaum, C. (2010). Wordnet. *Theory and applications of ontology: Computer applications* (pp. 231–243). Springer.

Gentile, C., & Warmuth, M. K. (1998). Linear hinge loss and average margin. *Advances in neural information processing systems*, *11*.

Gilda, S. (2017). Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, 110–115. https://doi.org/10.1109/SCORED.2017.8305411

Glazkova, A., Glazkov, M., & Trifonov, T. (2020). G2tmn at constraint@ aaai2021: Exploiting ct-bert and ensembling learning for covid-19 fake news detection. *arXiv preprint arXiv:2012.11967*.

Gupta, K., & Potika, K. (2021). Fake news analysis and graph classification on a covid-19 twitter dataset. *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, 60–68. https://doi.org/10.1109/BigDataService52369.2021.00013

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2009). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

Hamid, A., Sheikh, N., Said, N., Ahmad, K., Gul, A., Hassan, L., & Al-Fuqaha, A. I. (2020). Fake news detection in social media using graph neural networks and NLP techniques: A COVID-19 use-case. *CoRR*, *abs/2012.07517*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hörtenhuemer, C., & Zangerle, E. (2020). A Multi-Aspect Classification Ensemble Approach for Profiling Fake News Spreaders on Twitter—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., & Zhou, M. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 754–763. https://doi.org/10.18653/v1/2021.acl-long.62

Jiang, T., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). A novel stacking approach for accurate detection of fake news. *IEEE Access*, *9*, 22626–22639. https://doi.org/10.1109/ACCESS.2021.3056079

Kadam, A. B., & Atre, S. R. (2020). Negative impact of social media panic during the COVID-19 outbreak in India [taaa057]. *Journal of Travel Medicine*, *27*(3). https://doi.org/10.1093/jtm/taaa057

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). Deepfake: Improving fake news detection using tensor decomposition-based deep neural network. *The Journal of Supercomputing*, *77*(2), 1015–1037.

Kazemi, S. M., & Poole, D. (2018). Simple embedding for link prediction in knowledge graphs. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, december 3-8, 2018, montréal, canada* (pp. 4289–4300).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings*.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Koloski, B., Škrlj, B., & Robnik-Šikonja, M. (2020). Knowledge graph-based document embedding enrichment.

Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, *496*, 208–226. https://doi.org/https://doi.org/10.1016/j.neucom.2022.01.096

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 177–188.

Koloski, B., Pollak, S., & Škrlj, B. (2020). Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2011). Erratum: Estimating mutual information [phys. rev. e 69, 066138 (2004)]. *Physical Review E*, *83*(1), 019903.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 63–70. https://doi.org/10.3115/1118108.1118117

Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

Martinc, M., Skrlj, B., & Pollak, S. (2018). Multilingual gender classification with multi-view deep learning: Notebook for PAN at CLEF 2018. In L. Cappellato, N. Ferro, J. Nie, & L. Soulier (Eds.), *Working notes of CLEF 2018 - conference and labs of the evaluation forum, avignon, france, september 10-14, 2018*. CEUR-WS.org.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Moiseev, F., Dong, Z., Alfonseca, E., & Jaggi, M. (2022). Skill: Structured knowledge infusion for large language models. https://doi.org/10.48550/ARXIV.2205.08184

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: Anatural language processing model to analyse covid-19 content on twitter.

Murrieta Bello, H., Heilmann, L., & Ronan, E. (2020). Detecting Fake News Spreaders with Behavioural, Lexical and Psycholinguistic Features—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Nguyen, V.-H., Sugiyama, K., Nakov, P., & Kan, M.-Y. (2020). Fang: Leveraging social context for fake news detection using graph representation. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1165–1174. https://doi.org/10.1145/3340531.3412046

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., & Gipp, B. (2019). Enriching BERT with knowledge graph embeddings for document classification. *Proceedings of the GermEval Workshop 2019 – Shared Task on the Hierarchical Classification of Blurbs*.

Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., PYKL, S., Das, A., Ekbal, A., Akhtar, M. S., & Chakraborty, T. (2021). Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT)*.

Patwa, P., Sharma, S., PYKL, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2020). Fighting an infodemic: Covid-19 fake news dataset. *arXiv preprint arXiv:2011.03327*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Perdih, T. S., Lavrač, N., & Škrlj, B. (2021). Semantic reasoning from model-agnostic explanations. *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 000105–000110. https://doi.org/10.1109/SAMI50585.2021.9378668

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. https://doi.org/10.18653/v1/P18-1022

Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health*, *17*(7). https://doi.org/10.3390/ijerph17072430

Rangel, F., Giachanou, A., Ghanem, B., & Rosso, P. (2020). Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings
Conference and Labs of the Evaluation Forum (CLEF 2020).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), 3982–3992. https://doi.org/10.18653/v1/D19-1410

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, *abs/1910.01108*.

Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, *46*(2), 499–510. https://doi.org/10.1162/coli_a_00380

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36.

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. *Disinformation, misinformation, and fake news in social media* (pp. 1–19). Springer.

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., & Pollak, S. (2021). Tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*, *65*, 101104. https://doi.org/https://doi.org/10.1016/j.csl.2020.101104

Song, X., Wang, H., Zeng, K., Liu, Y., & Zhou, B. (2021). Katgcn: Knowledge-aware attention based temporal graph convolutional network for multi-event prediction.

Sun, Z., Deng, Z., Nie, J., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Tavchioski, I., Koloski, B., Škrlj, B., & Pollak, S. (2022). E8-IJS@ LT-EDI-ACL2022-BERT, AutoML and Knowledge-graph backed Detection of Depression. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 251–257.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In M. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33nd international conference on machine learning, ICML 2016, new york city, ny, usa, june 19-24, 2016* (pp. 2071–2080). JMLR.org.

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B. .-. (2020). Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, *8*, 156695–156706. https://doi.org/10.1109/ACCESS.2020.3019735

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 5998–6008).

Vovk, V. (2015). The fundamental nature of the log loss function. *Fields of logic and computation ii* (pp. 307–318). Springer.

Vrandečić, D., & Krötzsch, M. (2014). WikiData: A free collaborative knowledgebase. *Commun. ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. https://doi.org/10.18653/v1/P17-2067

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph and text jointly embedding. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1591–1601. https://doi.org/10.3115/v1/D14-1167

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings*.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: reasoning with language models and knowledge graphs for question answering. *CoRR*, *abs/2104.06378*.

Zeng, J., Zhang, Y., & Ma, X. (2021). Fake news detection for epidemic emergencies via deep correlations between text and images. *Sustainable Cities and Society*, *66*, 102652. https://doi.org/https://doi.org/10.1016/j.scs.2020.102652

Zhang, J., Dong, B., & Yu, P. S. (2020). Fakedetector: Effective fake news detection with deep diffusive neural network. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1826–1829. https://doi.org/10.1109/ICDE48307.2020.00180

Zhang, S., Tay, Y., Yao, L., & Liu, Q. (2019). Quaternion knowledge graph embeddings. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 2731–2741).

Zhu, Z., Xu, S., Tang, J., & Qu, M. (2019). Graphvite: A high-performance CPU-GPU hybrid system for node embedding. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, & L. Zia (Eds.), *The world wide web conference, WWW 2019, san francisco, ca, usa, may 13-17, 2019* (pp. 2494–2504). ACM. https://doi.org/10.1145/3308558.3313508

Zosa, E., Boros, E., Koloski, B., & Pivovarova, L. (2022). EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Bibliography

Journal Articles

Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, *496*, 208–226. https://doi.org/https://doi.org/10.1016/j.neucom.2022.01.096

Conference Paper

Koloski, B., Montariol, S., Purver, M., & Pollak, S. (2022). Knowledge informed sustainability detection from short financial texts.

Koloski, B., Pollak, S., & Skrlj, B. (2020). Know your neighbors: Efficient author profiling via follower tweets. *CLEF (Working Notes)*.

Koloski, B., Pollak, S., Škrlj, B., & Martinc, M. (2021a). Extending neural keyword extraction with TF-IDF tagset matching. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 22–29.

Koloski, B., Pollak, S., Škrlj, B., & Martinc, M. (2021b). Keyword extraction datasets for croatian, estonian, latvian and russian 1.0.

Koloski, B., Pollak, S., Škrlj, B., & Martinc, M. (2022). Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? *arXiv preprint arXiv:2202.06650*.

Koloski, B., Škrlj, B., & Robnik-Šikonja, M. (2020). Knowledge graph-based document embedding enrichment.

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 177–188.

Pollak, S., Robnik-Šikonja, M., Purver, M., Boggia, M., Shekhar, R., Pranjić, M., Salmela, S., Krustok, I., Paju, T., Linden, C.-G., et al. (2021). Embeddia tools, datasets and challenges: Resources and hackathon contributions.

Repar, A., Koloski, B., Ulcar, M., & Pollak, S. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. *LREC 2022 Workshop Language Resources and Evaluation Conference 25 June 2022*, 61.

Tavchioski, I., Koloski, B., Škrlj, B., & Pollak, S. (2021). Multi-label classification of COVID-19-related articles with an autoML approach.

Tavchioski, I., Koloski, B., Škrlj, B., & Pollak, S. (2022). E8-IJS@ LT-EDI-ACL2022-BERT, AutoML and Knowledge-graph backed Detection of Depression. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 251–257.

Zosa, E., Boros, E., Koloski, B., & Pivovarova, L. (2022). EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity.

Biography

Boshko Koloski was born on 8 January 1999 in Bitola, Macedonia. He completed his primary and secondary education in Bitola. In 2017, he enrolled the study Computer Science at the Faculty of Computer and Information Science of the University in Ljubljana. In 2020, he finished his undergraduate studies after defending his thesis "Knowledge graph-based document embedding enrichment" under the supervision of Prof. Dr. Marko Robnik-Šikonja and co-supervision Dr. Blaž Škrlj.

In the same year, he enrolled in the MSc program of Information and Communication Technologies at the Jožef Stefan International Postgraduate School in Ljubljana, where Assist. Prof. Senja Pollak and Dr. Blaž Škrlj act as co-supervisors.

In his studies, he focused on learning from heterogeneous data sources, ranging from texts to graphs. His work led to multiple published conference papers and one Neurocomputing journal paper. He also presented his work at several international conferences. He also collaborated in several projects including the H2020 EMBEDDIA project.