EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D4.7: Final cross-lingual news viewpoints identification technology (T4.3)

Executive summary

Task T4.3 (Cross-lingual identification of viewpoints and sentiment in news reporting) addresses monolingual and cross-lingual identification of viewpoints and sentiment in news reporting. This deliverable presents final viewpoints and sentiment detection technology developed. When it comes to viewpoint detection, we present several newly developed approaches that were employed for diachronic analysis (i.e. by detecting word usage changes over time) and ideological analysis (i.e. by detecting distinct word usages across media with different political background) in a multilingual setting. Next, we describe a novel methodology for news sentiment analysis based on sentiment-enrichment of BERT-based models and show that significant gains can be obtained by employing this methodology. Finally, an entire chapter is dedicated to our novel work on fake news detection, where we present approaches for classification of fake news and identification of fake news spreaders.

Partner in charge: JSI

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D4.7**
Deliverable full title:	**Final cross-lingual news viewpoints identification technology**
Deliverable short title:	**Cross-lingual viewpoints identification**
Document identifier:	**EMBEDDIA-D47-CrosslingualViewpointsIdentification-T43-submitted**
Lead partner short name:	**JSI**
Report version:	**submitted**
Report submission date:	**31/10/2021**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Matej Martinc (JSI)**
Co-author(s):	**Andraž Pelicon (JSI), Boshko Koloski (JSI), Senja Pollak (JSI), Lidia Pivovarova (UH)**
Status:	**__ draft, __ final, _x_ submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
06/09/2021	v0.1	Matej Martinc (JSI)	Report structure.
10/09/2021	v0.2	Matej Martinc (JSI)	Written section about diachronic change and viewpoint detection.
27/09/2021	v0.3	Andraž Pelicon (JSI)	Written section about sentiment analysis.
28/09/2021	v0.4	Boshko Koloski (JSI)	Written section about fake news detection.
29/09/2021	v0.5	Matej Martinc (JSI)	Written introduction and conclusion sections, proofreading all the content.
30/09/2021	v0.6	Lidia Pivovarova (UH)	Added content to diachronic change detection section.
04/10/2021	v0.7	Senja Pollak (JSI)	Added content about viewpoint detection on Slovenian COVID-19 corpus.
11/10/2021	v0.8	Jose G Moreno (ULR)	Internal review.
13/10/2021	v0.9	Saturnino Luz (UE)	Internal review.
22/10/2021	v0.10	Matej Martinc (JSI)	Addressed internal review comments.
22/10/2021	v0.11	Nada Lavrač (JSI)	Quality control.
26/10/2021	final	Matej Martinc (JSI)	Report finalized.
29/10/2021	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

1. Introduction **5**

2. Diachronic and ideological viewpoint analysis **5**

2.1 Background and related work 6

2.2 Methodology 8
2.2.1 Embeddings generation 9
2.2.2 Clustering 10
2.2.3 Change detection and interpretation 10
2.2.4 Grammatical profiling for semantic change detection 11

2.3 Diachronic viewpoint experiments on SemEval and Aylien corpora 12
2.3.1 Use case: Aylien COVID-19 corpus 13

2.4 Analysis of discourse dynamics with topic modelling techniques 15

2.5 Ideological viewpoint detection experiments on the LGBTIQ+ news corpus 16

2.6 Ideological viewpoint detection experiments on the Slovenian COVID-19 news corpus 18

3. Sentiment analysis **22**

3.1 Background and related work 23

3.2 Methodology 23

3.3 Experiments 25
3.3.1 Use case: Sentiment analysis on the LGBTIQ+ corpus 25

4. Fake news identification **26**

4.1 Data sets 26

4.2 Document representations considered 27

4.3 Knowledge graph-based document representations 28

4.4 Construction of the final representation 29

4.5 Classification models considered 30

4.6 Experiments 31
4.6.1 Task 1: LIAR 31
4.6.2 Task 2: FakeNewsNet 32
4.6.3 Task 3: PAN2020 32
4.6.4 Task 4: COVID-19 33

5. Conclusions and further work **33**

6. Associated outputs **34**

Appendix A: Scalable and Interpretable Semantic Change Detection **43**

Appendix B: EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+ 54
Appendix C: Zero-Shot Learning for Cross-Lingual News Sentiment Classification 60
Appendix D: COVID-19 V slovenskih spletnih medijih : analiza s pomočjo računalniške obdelave jezika 81
Appendix E: Knowledge Graph informed Fake News Classification via Heterogeneous Representation Ensembles 92
Appendix F: Grammatical Profiling for Semantic Change Detection 144
Appendix G: Benchmarks for Unsupervised Discourse Change Detection 156
Appendix H: Topic modelling discourse dynamics in historical newspapers 168

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
COHA	Corpus of Historical American
DoA	Description of Action
EC	European Commission
ELMo	Embeddings from Language Models
ExM	Ekspress Meedia
GA	Grant Agreement
JSD	Jensen-Shannon divergence
JSI	Jožef Stefan Institute
UH	University of Helsinki
LSTM	Long short-term memory
NBM	Naive Bayes Multinominal
NE	Named entity
NLP	Natural Language Processing
RAM	Random Access Memory
OP	Orthogonal Procrustes
LDA	Latent Dirichlet Allocation
DTM	Dynamic Topic Modelling
LSA	Latent Semantic Analysis
WD	Wasserstein Distance
CD	Cosine Distance
SVD	Singular Value Decomposition
SemEval	Semantic Evaluation
SGNS	Skip-gram negative sampling
KG	Knowledge graph
KB	Knowledge base
T	Task
TF-IDF	Term Frequency–Inverse Document Frequency
WP	Work Package

1 Introduction

The overall objective of WP4, named *Cross-lingual content analysis*, is to facilitate the analysis of news content across different languages, empowering news media consumers, researchers and news media professionals. This report, entitled *Final cross-lingual news viewpoints identification technology* is a result of the activities performed in Task T4.3, and focuses on presenting the final technology for identifying viewpoints and sentiment that can be employed in a multilingual setting. Specifically, this deliverable describes the final results achieved in viewpoint identification and sentiment analysis (SA) within Task T4.3 (Cross-lingual identification of viewpoints and sentiment in news reporting), which started in M7 and lasts until M33. On top of that, we also report about our work on the topic of fake news detection, a problem which has gained a lot in significance in the recent years.

The main contributions presented in this deliverable are the following:

- **Diachronic and ideological viewpoint detection technology**: Building on our previous work on this topic, which was presented in deliverable D4.4, we present several new methods for news analysis that can be employed for diachronic and ideological viewpoint detection in a multilingual setting. We show that the proposed methods are scalable and interpretable, and also offer better performance than previous interpretable methods. The work on this topic was conducted in collaboration of UH and JSI, and is published in papers by Montariol et al. (2021); Martinc et al. (2021); Giulianelli et al. (2021); Duong et al. (2021); Marjanen et al. (2021); Pollak et al. (2021), see appendices A, B, D, F, G and H, respectively.
- **Sentiment analysis methods**: We present a novel approach for news sentiment analysis, which is based on our previous work on this topic described in deliverable D4.4. The approach is based on a novel technique for injecting sentiment knowledge into BERT models through intermediate training. The method was tested on a large manually labeled Slovenian news dataset and the results indicate that substantial gains in performance can be obtained by employing this tactic. The trained model was also used for sentiment detection in the viewpoint study of the LGBTIQ+ corpus, where we compared sentiment distribution of articles of media with distinct ideological backgrounds. The work on this topic was presented in a journal publication (Pelicon et al., 2020) and a conference publication (Martinc et al., 2021), see appendices C and B, respectively.
- **Fake news detection**: We present our recent approaches for tackling the fake news problem, which can be employed for classification of fake news and identification of fake news spreaders. The approaches are based on employing multi-modal representations constructed from text and knowledge graph input. The proposed approaches were tested on several shared tasks with good results. The work on this topic was published in paper (Koloski, Stepišnik-Perdih, et al., 2021), see appendix E.

The deliverable is organised as follows. We start with the presentation of diachronic and ideological viewpoint analysis, which is covered in Section 2. In Section 3 we describe the novel approach for sentiment analysis using BERT sentiment enrichment and Section 4 describes our work on the topic of fake news detection. The report finishes with conclusions and plans for further work (Section 5), and with the associated outputs of the work done within T4.3 (Section 6).

2 Diachronic and ideological viewpoint analysis

In the previous deliverable for T4.3 we have presented two distinct methods for diachronic viewpoint detection, which considered language evolution as a proxy for diachronic news analysis and looked for changes in usage of specific words[1] that would indicate temporal viewpoint change. The first method relied on averaging of contextual BERT embeddings (Martinc, Novak, & Pollak, 2020) in order to derive

[1] Note that we refer to all types of language evolution—short- or long-term, with or without meaning change—as word usage change, a broad category that includes semantic change, but also any shifts in the context in which a word appears.

a specific temporal representation for a specific word. The second method relied on clustering of contextual embeddings (Martinc, Montariol, et al., 2020b,a): if clusters, which in theory capture distinct word usages, are distributed differently across time periods, it indicates a possible change in word's context or even loss or gain of a word sense. Thus, the cluster-based approach offers a more intuitive interpretation of word usage change than alternative methods, which look at the neighborhood of a word in each time period to interpret the change (Gonen et al., 2020; Martinc, Novak, & Pollak, 2020) and ignore the fact that a word can have more than one meaning. The main limitation of the cluster-based methods is the scalability in terms of memory consumption and time: clustering is applied to each word in the corpus separately and all occurrences of a word need to be aggregated into clusters. For large corpora with large vocabularies, where some words can appear millions of times, the use of these methods is severely limited.

To avoid the scalability issue, cluster-based methods are generally applied to a small set of less than a hundred manually pre-selected words (Giulianelli et al., 2020; Martinc, Montariol, et al., 2020a). This drastically limits the application of the methods in scenarios such as identification of the most changed words in a large corpus or measuring of usage change of extremely frequent words, since clustering of all of word's contextual embeddings requires large computational resources. One way to solve the scalability problem using contextual embeddings is to *average* a set of contextual representations for each word into a single static representation (Martinc, Novak, & Pollak, 2020). Averaging, while scalable, loses a lot on the interpretability aspect, since word usages are merged into a single representation.

The method we propose in this final deliverable for T4.3 and was also presented in paper (Montariol et al., 2021), which is attached as Appendix A, tries to alleviate deficiencies of previous methods and tackles scalability and interpretability at the same time. To put it differently, we propose a *scalable* method for contextual embeddings clustering that generates interpretable representations and outperforms other cluster-based methods. Additionally, we also propose an *interpretation pipeline* that automatically labels word senses, allowing a domain expert to find the most changing concepts and to understand *how* those changes happened.

The practical abilities of our method are first demonstrated on a large corpus of news articles related to COVID-19, the Aylien Coronavirus News Dataset[2]. We compute the degree of usage change of almost 8,000 words, i.e. all words that appear more than 50 times in every time slice of the corpus, in the collection of about half a million articles in order to find the most changing words and interpret their drift.

In the next step, we deploy the method on non-diachronic tasks of determining viewpoints of different Slovenian media providers. More specifically, we created two corpora, a corpus of LGBTIQ+ related Slovenian news and a corpus of COVID-19 related Slovenian news, and conducted an automatic analysis of its content, trying to identify the words that are used differently in different news sources and would indicate the difference in the prevailing discourse on the topics of LGBTIQ+ and COVID-19 in the specific liberal and conservative media.

2.1 Background and related work

Among distinct viewpoint detection tasks, the analysis of diachronic change is the one that gained most traction recently. Diachronic word embedding models have undergone a surge of interest in the last two years with the successive publications of three articles dedicated to a literature review of the domain (Kutuzov et al., 2018; Tahmasebi et al., 2018; Tang, 2018). Most approaches build static embedding models for each time slice of the corpus and then make these representations comparable by either employing *incremental updating* (Kim et al., 2014) or *vector space alignment* (Hamilton et al., 2016b).

In the former approach, an embedding matrix is trained on the first time slice of the corpus and updated at each successive time slice using the previous matrix for initialisation. In the latter approach, an em-

[2] https://blog.aylien.com/free-coronavirus-news-dataset/

bedding space is trained on each time slice independently and an alignment is performed by optimising a geometric transformation. The alignment method has proved superior on a set of synthetic semantic drifts (Shoemark et al., 2019) and has been extensively used (Hamilton et al., 2016b; Dubossarsky et al., 2017) and improved (Dubossarsky et al., 2019) in the literature. The recent SemEval Task on unsupervised lexical semantic change detection has shown that this method is most stable and yields the best averaged performance across four SemEval corpora Schlechtweg et al. (2020). Yet another approach (Hamilton et al., 2016a; Yin et al., 2018) is based on comparison of neighbors of a target word in different time periods. This approach has been recently used to tackle the scalability problem (Gonen et al., 2020).

In all these methods above, each word has only one representation within a time slice, which limits the interpretability of these techniques. The recent rise of contextual embeddings such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) introduced significant changes to word representations. These models allow each word occurrence to have a vector representation that depends on its context. When pre-trained on large datasets, they improve the state-of-the-art on numerous NLP tasks. Contextual embeddings can be used for usage change detection by aggregating the information from the set of token embeddings. This can be done either through averaging of all vectors within a time slice and then computing averaged vector similarity (Martinc, Novak, & Pollak, 2020), by computing a pairwise distance between vectors from different time slices (Kutuzov & Giulianelli, 2020), or by clustering all token representations to approximate its set of senses (Giulianelli et al., 2020). The analysis in this paper derives from this last set of methods, which demonstrate a higher performance than static embeddings methods at least on some datasets (Martinc, Montariol, et al., 2020a).

Automatic semantic shift detection methods have been used for text stream monitoring tasks, such as event detection (Kutuzov et al., 2017), viewpoint analysis (Azarbonyad et al., 2017) or monitoring of rapid discourse changes during crisis events (Stewart et al., 2017). None of these applications use clustering techniques and, as far as we are aware, only Martinc, Novak, & Pollak (2020) uses contextual embeddings for news stream analysis. When it comes to methods employed for a specific task of automatic analysis of the LGBTIQ+ topic, which we tackle in this deliverable, most recent approaches rely on static embeddings. Hamilton et al. (2016b) employed embeddings to research how words (among them also word *gay*) change meaning through time. They built static embedding models for each time slice of the corpus and then make these representations comparable by employing *vector space alignment* by optimising a geometric transformation. This research was recently expanded by Shi & Lei (2020), who employed static embeddings to explore semantic shifts of six descriptive LGBTIQ+ words from the 1860s to the 2000s: *homosexual, lesbian, gay, bisexual, transgender*, and *queer*.

The related work on quantitative content analysis of news related to LGBTIQ+ nevertheless indicates that the employment of the proposed method for word usage change detection could be used for detection of discursive differences between media with different political orientation. Previous analysis showed that distinctions can be drawn between those media articles that express positive, neutral or negative stance towards same-sex marriage. Those media articles that express positive stance are grounded in human rights/civil equality discourses and access to benefits (Zheng & Chan, 2020; Colistra & Johnson, 2019; Paterson & Coffey-Glover, 2018), and frame marriage equality as an inevitable path towards equality, as a civil right issue that would reduce existing prejudices and discrimination, and protect threatened LGBTIQ+ minority (Zheng & Chan, 2020).

For media articles that express negative stance towards marriage equality, distinctive discursive elements are present, such as “equal, but separate” (marriage equality should be implemented, but differentiating labels should be kept in the name of protecting the institute of marriage) (Kania, 2020; Zheng & Chan, 2020; Paterson & Coffey-Glover, 2018), and reference procreation/welfare of children (Kania, 2020; Zheng & Chan, 2020), public objection (Kania, 2020) and church – state opposition (Paterson & Coffey-Glover, 2018). The related work also shows that the differences between "liberal" and "conservative" arguments are not emphasised, mostly because both sides refer to each other’s arguments, if only to negate them; yet, political orientation can be identified through the tone of the article (Zheng & Chan, 2020).

The main motivation for the proposed research on the topic of viewpoint detection are scalability and interpretability issues of previous methods for word usage change detection. The ones using contextual embeddings are either interpretable but unscalable (Giulianelli et al., 2020; Martinc, Montariol, et al., 2020a) or scalable but uninterpretable (Martinc, Novak, & Pollak, 2020). The scalability issues of interpretable methods can be divided into two problems.

Memory consumption: Giulianelli et al. (2020) and Martinc, Montariol, et al. (2020a) apply clustering on all embeddings of each target word. This procedure becomes unfeasible for large sets of target words or if the embeddings need to be generated on a large corpus, since too many embeddings need to be saved into memory for further processing. To give an example, single-precision floating-point in Python requires 4 bytes of memory. Each contextual embedding contains 768 floats (Devlin et al., 2019), leading each embedding to occupy 3072 bytes[3]. To use the previous methods on the Aylien Coronavirus News Dataset[4], which contains 250M tokens, about 768 Gb RAM would be necessary to store the embeddings for the entire corpus. If we limit our vocabulary to the 7,651 words that appear at least 50 times in every time slice and remove the stopwords (as we do in this work), we still need to generate contextual embeddings for 120M tokens, which is about 369 Gb of RAM.

Complexity of clustering algorithms: For the complexity analyses, we denote by d the dimension of the embedding, k is the number of clusters and n is the number of contextual embeddings, i.e. the number of word occurrences in the corpus. The time complexity of the affinity propagation algorithm (the best performing algorithm according to Martinc, Montariol, et al. (2020a)) is $O(n^2td)$, with t being the predefined maximum number of iterations of the data point message exchange. The time complexity of the simpler k-means algorithm[5] can be stated as $O(tknd)$, where t is the number of iterations of Lloyd's algorithm (Lloyd, 1982). As an example, consider the word *coronavirus*, which appears in the Aylien corpus about 1,2M times. For k-means with $k = 5$ and a maximal number of iterations set to 300 (the Scikit library default), about $300*5*1,300,000*768 \approx 1.5 \times 10^{12}$ operations are conducted for the clustering. With affinity propagation with the maximum number of iterations set to 200 (the default), clustering of the word *coronavirus* would require $1,300,000^2*200*768 \approx 2.6 \times 10^{17}$ operations, which is impossible to conduct in a reasonable amount of time on a high end desktop computer.

Scalable methods that employ contextual embeddings on the other hand have interpretability limitations. The averaging approach (Martinc, Novak, & Pollak, 2020) eliminates the scalability problems: token embeddings for each word are not collected in a list but summed together in an element-wise fashion, which means that only 768 floats need to be saved for each word in the vocabulary. The averaged word representation is obtained for each time slice by dividing the sum by the word count. A single embedding per word is saved, leading to only 23.5 Mb of RAM required to store the embeddings for 7,651 words. These representations loose on the interpretability aspect, since all word usages are merged into a single averaged representation. It makes the method inappropriate for some tasks such as automatic labelling of word senses, and in some cases affects the overall performance of the method (Martinc, Montariol, et al., 2020a).

2.2 Methodology

Our word usage change/viewpoint detection pipeline follows the procedure proposed in the previous work (Martinc, Montariol, et al., 2020a; Giulianelli et al., 2020): for each word, we generate a set of contextual embeddings using BERT (Devlin et al., 2019). These representations are clustered using k-means or affinity propagation and the derived cluster distributions are compared across time/media-source slices by either using Jensen-Shannon divergence (JSD) Lin (2006) or the Wasserstein distance

[3] If we ignore the additional memory of a Python container—e.g., a Numpy list or a Pytorch tensor—required for storing this data.

[4] `https://blog.aylien.com/free-coronavirus-news-dataset/`

[5] Here we are referring to the Scikit implementation of the algorithm employed in this work: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

(WD) (Solomon, 2018). Finally, words are ranked according to the distance measure, assuming that the ranking resembles a relative degree of usage shift.

The primary contributions of this work lay in the embedding generation step, which improves the scalability of the method, and in leveraging WD to compute the distance between clusters, which improves performance of the method. We also propose post-processing steps, which domain experts could use for the interpretation of results. We now describe the pipeline in more details.

2.2.1 Embeddings generation

We use a pre-trained BERT model for each language of the evaluation corpora.[6] All models have 12 attention layers and a hidden layer of size 768. We fine-tune them for domain adaptation on each corpus as a masked language model for 5 epochs. Then, we extract token embeddings from the fine-tuned models. Each corpus is split into time/media-source slices. The models are fed 256 tokens long sequences in batches of 16 sequences at once. We generate sequence embeddings by summing the last four encoder output layers of BERT, following Devlin et al. (2019). Next, we split each sequence into 256 sub-parts to obtain a separate contextual embedding of size 768 for each token. Since one token does not necessarily correspond to one word due to byte-pair tokenization, we average embeddings for each byte-pair token constituting a word to obtain embeddings for each occurrence of a word.

Next, after obtaining a contextual embedding vector for each target word in a specific sequence, we decide whether this vector should be *saved* to the list or *merged* with one of the previously obtained vectors for the same word in the same time/media-source slice. To improve the scalability, we limit the number of contextual embeddings that are kept in the memory for a given word and time/media-source slice to a predefined threshold. The threshold of 200 was chosen empirically from a set of threshold candidates (20, 50, 100, 200, 500) and offers a reasonable compromise between scalability and performance. The new vector is merged if it is too similar—i.e. a duplicate or a near-duplicate—to one of the saved vectors or if the list already contains a predefined maximum number of vectors (200 in our case).

More formally, we add the new embedding e_{new} to the list of word embeddings $L = \{e_i, ..., e_n\}$ if:

$$|L| < 200 \quad \& \quad \forall e_i \in L\colon s(e_{\text{new}}, e_i) < 1 - \varepsilon$$

where s is the cosine similarity and ε is a threshold set to 0.01.

If $|L| \geq 200$ or if any vector in the list L is a near duplicate to e_{new}, we find a vector e_m in the list which is the closest to e_{new} in terms of cosine similarity:

$$e_m = \arg\max_{e_i \in L} s(e_i, e_{\text{new}})$$

This element e_m is then modified by summing it with e_{new}:

$$e_m \leftarrow e_m + e_{\text{new}}$$

The number of summed-up elements for each of the 200 groups in the list is stored besides their summed-up representations. Once the model has been fed with all the sequences in the time/media-source slice, the final summed-up vector is divided by this number to obtain an averaged embedding.

[6] For German: bert-base-german-cased (`https://deepset.ai/german-bert`), for English: bert-base-uncased model, for Latin: bert-base-multilingual-uncased model from the `huggingface` library, for Swedish: bert-base-swedish-uncased (`https://github.com/af-ai-center/SweBERT`), for Slovenian: EMBEDDIA/crosloengual-bert (`https://www.clarin.si/repository/xmlui/handle/11356/1330`) (Ulčar & Robnik-Šikonja, 2020).

By having only 200 merged word embeddings per word per time/media-source slice, and by limiting the vocabulary of the corpus to 7,651 target words, we require up to 4.7 Gb of space for each time/media-source slice, no matter the size of the corpus. While this is still 200 times more space than if the averaging method was used (Martinc, Novak, & Pollak, 2020), the conducted experiments show that the proposed method nevertheless keeps the bulk of the interpretability of the less scalable method proposed by Giulianelli et al. (2020), and offers competitive performance on several corpora.

2.2.2 Clustering

After collecting 200 vectors for each word in each time/media-source slice, we conduct clustering on these lists to extract the usage distribution of the word at each period. Clustering for a given word is performed on the set of all vectors from all time/media-source slices jointly.

We use two clustering methods previously applied for this task, namely k-means used in Giulianelli et al. (2020) and affinity propagation in Martinc, Montariol, et al. (2020a). The main strength of affinity propagation is that the number of clusters is not defined in advance but inferred during training. The clustering is usually skewed: a limited number of large clusters is accompanied with many clusters consisting of only a couple of instances. Thus, affinity propagation allows to pick out the core senses of a word. K-means tends to produce more even clusters. Appearance of small clusters that contain only few instances and do not represent a specific sense or usage of the word is nevertheless relatively common, since BERT is sensitive to syntax and pragmatics, which are not necessarily relevant for usage change detection. Another limitation of the k-means algorithm is that the number of clusters needs to be set in advance. This means that if the number of actual word usages is smaller than a predefined number of clusters, k-means will generate more than one cluster for each word usage.

To compensate for these deficiencies, we propose an additional *filtering and merging* step. A cluster is considered to be a legitimate representation of a usage of the word, if it contains at least 10 instances[7]. We compute the average embedding inside each cluster, and measure the cosine distance (1 - cosine similarity) between the average embeddings in each pair of legitimate clusters for a given word. If the distance between two clusters is smaller than a threshold, the clusters are merged. The threshold is defined as $avg_{cd} - 2 * std_{cd}$, where avg_{cd} is the average pairwise cosine distance between all legitimate clusters and std_{cd} is the standard deviation of that distance. This merging procedure is applied recursively until the minimum distance between the two closest clusters is larger than the threshold. After that, the merging procedure is applied to illegitimate clusters (that contain less than 10 instances), using the same threshold. Illegitimate clusters could be added into one of the legitimate clusters or merged together to form a legitimate cluster with more than 10 instances. If there is no cluster that is close enough to be merged with, the illegitimate cluster is removed.

2.2.3 Change detection and interpretation

After the clustering procedure described above, for each word in each time/media-source slice, we extract its cluster distribution and normalise it by the word frequency in the slice. Then target words are *ranked* according to the usage divergence between time/media-source slices, measured with the JSD or the WD[8]. If a ground-truth ranking exists, the method can be evaluated using the Spearman Rank Correlation to compare the true and the outputted ranking. In the exploratory scenario, the ranking is used to detect the most differently used words and then investigate the most unevenly distributed clusters over time/media-source for the interpretation of the change.

JSD has been used for semantic shift detection in several recent papers, e.g. Martinc, Montariol, et al.

[7]The threshold of 10 was derived from the procedure for manual labelling employed in the SemEval Task Schlechtweg et al. (2020), where a constraint was enforced that the specific sense is attested at least 5 times in a specific time period in order to contribute word senses. We set the overall threshold of 10, which roughly translates to 5 per time period, since all of our test corpora (besides Aylien, see Section 2.3.1) contain two time/media-source periods.

[8]Using the POT package `https://pythonot.github.io/`.

(2020a); Giulianelli et al. (2020); Kutuzov & Giulianelli (2020). Since we are the first that apply WD for this purpose, we describe it in more details. The motivation for using the WD (Solomon, 2018) is to take into account the position of the clusters in the semantic space when comparing them. The JSD leverages semantic information encoded in the embeddings indirectly, distilled into two time-specific cluster distributions that JSD receives as an input. In addition to cluster distributions, WD accesses characteristics of the semantic space explicitly, through a matrix of cluster averages (obtained by averaging embeddings in each cluster) of size $T \times k \times 768$, where k is a number of clusters, T is a number of time slices and 768 is the embedding dimension.

This setup is a classical problem that can be solved using optimal transport (Peyré et al., 2019). We denote with μ_1 and μ_2 the sets of k average embedding points in the two vector spaces, and with c_1 and c_2 the associated clusters distributions. Thus, c_1 and c_2 are histograms on the simplex (positive and sum to 1) that represent the weights of each embedding in the source (μ_1) and target (μ_2) distributions. The task is to quantify the effort of moving one unit of mass from μ_1 to μ_2 using a chosen cost function, in our case the cosine distance. It is solved by looking for the transport plan γ, which is the minimal effort required to reconfigure c_1's mass distribution into that of c_2. The WD is the sum of all travels that have to be made to solve the problem:

$$\mathrm{WD}(c_1, c_2) = \min_{\gamma} \sum_{i,j} \gamma_{i,j} M_{i,j}$$
$$\text{with } \gamma 1 = c_1;\ \gamma^{\mathsf{T}} 1 = c_2;\ \gamma \geq 0$$

Where $M \in \mathbb{R}^{+}_{m \times n}$ is the cost matrix defining the cost to move mass from μ_1 to μ_2. We use the cosine similarity s, with $M = 1 - s(\mu_1, \mu_2)$.

Interpretation. Once the most differently used words are detected, the next step is to understand *how* their usage differs between two time/media-source slices by interpreting their clusters of usages.

Cluster distributions can be used directly to identify the clusters that are unevenly distributed across a time/media-source dimension. However, a cluster itself may consist of several hundreds or thousands of word usages, i.e. sentences. Interpreting the underlying sense behind each cluster by manually looking at the sentences is time-consuming. To reduce human work, we extract the most discriminating words and bigrams for each cluster: by considering a cluster as a single document and all clusters as a corpus, we compute the term frequency - inverse document frequency (TF-IDF) score of each word and bigram in each cluster. The stopwords and the words appearing in more than 80% of the clusters are excluded to ensure that the selected keywords are the most discriminant. Thus, a ranked list of keywords for each cluster is obtained and top-ranked keywords are used for the interpretation of the cluster.

2.2.4 Grammatical profiling for semantic change detection

Though large embedding models proved to be efficient for semantic change detection, they obviously loose information on some language phenomena, especially grammatical. Thus, we conducted additional experiments aimed at investigation of how well semantic change detection could be performed with *grammatical profiling*, i.e. via morphological and syntactic features, without leveraging of any semantic information. This work was more theory-motivated, i.e. the goal was not to establish a new state-of-the-art but rather investigate complex relations between semantics and grammar. However, grammatical profiling yields surprisingly good evaluation scores across different languages and datasets, without any language-specific tuning, consistently outperforming count-based distributional baselines, which have access to lexical semantic data. For Latin, a language with rich morphology, grammar-based method even outperformed previous best result (Martinc, Montariol, et al., 2020b). This suggests that this approach could be especially fruitful for morphologically rich languages, such as the core Embeddia languages. However, practical implications of this work are yet to be established in further experiments.

More detailed description about the approach and the analysis are described in our paper Giulianelli et al. (2021), provided in Appendix F.

2.3 Diachronic viewpoint experiments on SemEval and Aylien corpora

We use six existing manually annotated datasets for evaluation. The first dataset, proposed by Gulordava & Baroni (2011), consists of 100 English words labelled by five annotators according to the level of semantic change between the 1960s and 1990s[9]. To build the dataset, the annotators evaluated semantic change using their intuition, without looking at the context. This procedure is problematic since an annotator may forget or not be aware of a particular sense of the word.

The organizers of the recent SemEval-2020 Task 1— Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020)—employed another approach: the annotators had to decide whether a pair of sentences from different time periods convey the same meaning of the word (Schlechtweg & Schulte im Walde, 2020). For each of the four languages—German, English, Latin and Swedish—senses were manually annotated by labeling word senses in a pair of sentences drawn from different time periods. All SemEval-2020 Task 1 corpora contain only two periods and the sentences are shuffled and lemmatized. The lexical semantic change score is defined as the difference between word sense frequency distributions in the two time periods and measured by the Jensen-Shannon Distance (Lin, 2006).

The DURel dataset (Schlechtweg et al., 2018) is composed of 22 German words, ranked by semantic change by five annotators between two time periods, 1750–1799 and 1850–1899. Similarly to SemEval, the ranking was build by evaluating the relatedness of pairs of sentences from two periods.

In order to conduct usage change detection on the target words proposed by Gulordava & Baroni (2011), we fine-tune the English BERT-base-uncased model and generate contextual embeddings on the Corpus of Historical American English (COHA)[10]. We only use data from the 1960s to the 1990s (1960s has around 2.8M and 1990s 3.3M words), to match the manually annotated data. For the SemEval Task 1 evaluation set, we fine-tune the BERT models and generate contextual embeddings on the four corpora provided by the organizers of the task, English (about 13.4M words), German (142M words), Swedish (182M words) and Latin (11.2M words). Finally, we fine-tune BERT and generate embeddings on the German DTA corpus (1750–1799 period has about 25M and 1850–1899 has 38M tokens)[11].

The results are shown in Table 1. We compare our scalable approach with the *non-scalable clustering* methods used by Giulianelli et al. (2020) and Martinc, Montariol, et al. (2020a). *Averaging* (Martinc, Novak, & Pollak, 2020) is the less interpretable method described in Section 2.1. *SGNS + OP + CD* (Schlechtweg et al., 2019) refers to the state-of-the-art semantic change detection method employing non-contextual word embeddings: the Skip-Gram with Negative Sampling (SGNS) model is trained on two periods independently and aligned using Orthogonal Procrustes (OP). Cosine Distance (CD) is used to compute the semantic change. The *Nearest Neighbors* method (Gonen et al., 2020) also uses SGNS embeddings. For each period, a word is represented by its top nearest neighbors (NN) according to CD. Semantic change is measured as the size of the intersection between the NN lists of two periods.

On average, the proposed scalable clustering with filtering and merging of clusters leads to a higher correlation with gold standard than the standard non-scalable clustering methods: the best method (affprop WD) achieving a Spearman correlation with the gold standard of 0.474 compared to the best non-scalable k-means 5 JSD achieving the Spearman correlation of 0.391. The method also outperforms averaging and NN, though it is outperformed by a large margin by the SGNS+OP+CD, achieving the score of 0.533.

[9] In order to make the proposed approach comparable to previous work, we remove four words that do not appear in the BERT vocabulary from the evaluation dataset, same as in Martinc, Montariol, et al. (2020a).

[10] https://www.english-corpora.org/coha/

[11] https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/durel/

Table 1: Spearman Rank Correlation between system output rankings and ground truth rankings for various datasets. "SE" stands for SemEval.

	COHA	SE English	SE Latin	SE German	SE Swedish	DURel	Avg. all
METHODS NOT USING CLUSTERING							
SGNS + OP + CD	0.347	0.321	0.372	**0.712**	**0.631**	**0.814**	**0.533**
Nearest Neighbors	0.310	0.150	0.273	0.627	0.404	0.590	0.392
Averaging	0.349	0.315	**0.496**	0.565	0.212	0.656	0.432
NON-SCALABLE CLUSTERING METHODS							
k-means 5 JSD	0.508	0.189	0.324	0.528	0.238	0.560	0.391
aff-prop JSD	**0.510**	0.313	0.467	0.436	-0.026	0.542	0.374
INTERPRETABLE SCALABLE METHODS							
Without filtering or merging of clusters							
k-means 5 JSD	0.430	0.316	0.358	0.508	0.073	0.658	0.390
aff-prop JSD	0.394	0.371	0.346	0.498	0.012	0.512	0.355
k-means 5 WD	0.372	0.360	0.450	0.514	0.316	0.607	0.437
aff-prop WD	0.369	**0.456**	0.397	0.421	0.264	0.484	0.399
With filtering and merging of clusters							
k-means 5 JSD	0.448	0.318	0.374	0.519	0.073	0.649	0.397
aff-prop JSD	0.403	0.348	0.408	0.583	0.018	0.712	0.412
k-means 5 WD	0.382	0.375	0.466	0.520	0.332	0.628	0.451
aff-prop WD	0.352	0.437	0.488	0.561	0.321	0.686	0.474

The best performing clustering algorithm differs for different datasets. On average, affinity propagation only outperforms k-means when filtering and merging of clusters is employed. The effect of the filtering on k-means is positive on average but the difference is thin, as the number of clusters is low.

WD leads to better results than JSD on most of the corpora where averaging outperforms clustering, the only exception is DURel. An extreme example is the Swedish SemEval dataset, where the clustering with JSD performs particularly poorly: using the WD, which takes into account the average embeddings on top of cluster distributions, greatly increases the correlation with the gold standard.

2.3.1 Use case: Aylien COVID-19 corpus

The combination of scalable clustering with the interpretation pipeline opens new opportunities for diachronic corpus exploration. In this section, we demonstrate how it could be used to analyze the Aylien Coronavirus News Dataset. The corpus contains about 500k news articles related to COVID-19 from January to April 2020[12], unevenly distributed over the months (160M words in March, 41M in February, 35M in April and 10M in January). We split the corpus into monthly chunks and apply our scalable word usage change detection method.

The scalable method allows to perform embeddings extraction and clustering for all words in the corpus. We extract the top words with the highest average WD between the successive months to conduct a deeper analysis. We exclude words that appear less than 50 times in each month to avoid spurious drifts due to words having too few occurrences in a time slice. However, some drifts due to corpus artefacts remain, in particular dates such as *'2019-20'*. Thus, words containing numbers and one-letter words are also removed.

In Table 2 we present the top 10 most drifting words extracted using k-means with k=5 and ranked according to the average WD across the four months[13]. Among them, the word *diamond* is related to the cruise ship "Diamond Princess", which suffered from an outbreak of COVID-19 and was quarantined for

[12]We used an older version of the corpus. Currently the data from May is also available.

[13]This is a rather arbitrary procedure: one can imagine that a domain expert would prefer a different frequency threshold or focus more on a given month. The most time-consuming part is embedding extraction. Once this is done, clustering and keyword extraction can be done as many times as necessary.

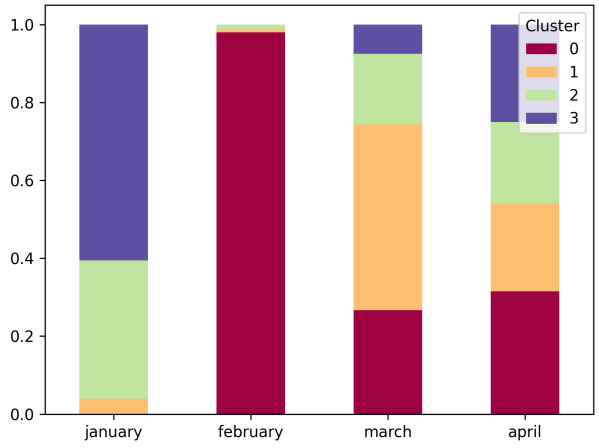
Table 2: Top 10 most changed words in the corpus according to a monthly-averaged WD of k-means ($k = 5$) cluster distributions.

1	diamond	6	tag
2	king	7	paramount
3	ash	8	lynch
4	palm	9	developers
5	fund	10	morris

several weeks. The word *king*, which is the second most changing word, is related to the King county, Washington, where the first confirmed COVID-19 related death in the USA appeared, and to the Netflix show "Tiger King", which was released in March. Thus, the primary context for this word changed several times, which is reflected in our results. Other words are mostly constituent words in named entities, related e.g., to an American Society of Hematology (ASH) Research Collaborative's Data Hub, which is capturing data on subjects tested positive for COVID-19.

The results suggest that the model does what it is meant to do: for most words in the list it is possible to find an explanation why its usage changed during the beginning of 2020. The list contains many proper names or proper name constituents, which could be either desirable or undesirable property, depending on research goals. Some work focuses specifically on proper names (Hennig & Wilson, 2020), since they could be a good proxy to shifts in socio-political situations. On the other hand, if the focus of the study are shifts in more abstract concepts, then proper names could be filtered out before the embedding generation stage by employing named entity recognition tools.

The interpretation pipeline, described in Section 2.2.3, is illustrated in Figures 1 and 2. We focus on two words, *diamond* and *strain*, to show the various phenomena that can be detected. *Diamond* is the top drifting word in the entire vocabulary (see Table 2); it can be both a common noun and an entity, inducing usage drift when the entity appears in the newspapers after events with high media coverage. *Strain* is the 38th word with the highest drift overall, and the 15th highest between February and March 2020. It has several different senses whose usage vary across time following the events in the news. We cluster their vector representations from the Aylien corpus using k-means with $k = 5$ and apply the cluster filtering and merging step. Then, using TF-IDF on unigrams and bigrams, we extract a set of keywords for each cluster to interpret the variations of their distribution.

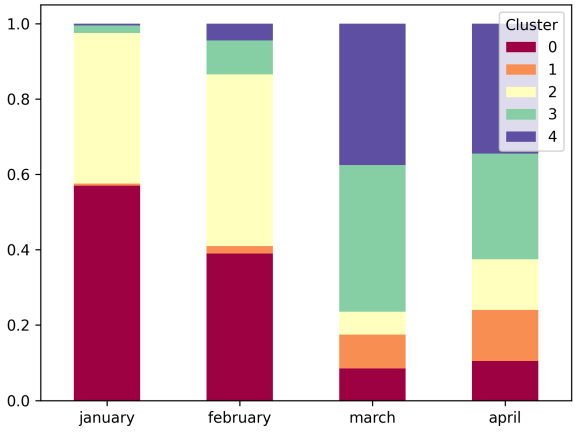


#	Keywords
0	diamond princess, cruise ship, princess cruise, japanese, tested positive, confirm, ship diamond
1	neil diamond, comic, sweet caroline, trump, song, diamond said, comic book,
2	diamond hill, hill capital, diamond jubilee, diamond mountain, league postponed, portfolio, athletics
3	diamond industry, black diamond, jewellery, hong kong, diamond ring, surat diamond, india

Figure 1: Cluster distributions per month and top keywords for each cluster for word *diamond*.

The keywords and cluster distributions for the word *diamond* can be found in Figure 1. One of the clusters was removed at the filtering step, as it had less than 10 embeddings inside, and no other cluster was close enough. A clear temporal tendency is visible from the cluster distribution in Figure 1: a new major usage appears in February, corresponding to the event of the quarantined cruise ship (Cluster

0); this association is revealed by the keywords for this cluster. Moreover, the WD between January and February, when the outbreak happened, is 0.337; it is also very high between February and March (0.342). It reflects the large gap between the cluster distributions, first with the appearance of Cluster 0 in February that made the other usages of the word *diamond* in the media almost disappear, and then the reappearance of other usages in March, when the situation around the cruise ship gradually normalized. Cluster 1, that appears in March, is related to Neil Diamond's coronavirus parody of the song "Sweet Caroline" which was shared mid-March on the social media platforms and received a lot of attention in the US. Cluster 3 is related to the diamond industry; it is much less discussed as soon as the pandemic breaks out in February. Finally, Cluster 2 deals with several topics: Diamond Hill Capital, a US investment company, and the Wanda Diamond League, an international track and field athletic competition which saw most of its meetings postponed because of the pandemic. This last cluster shows the limitations of our clustering: it is complex to identify and differentiate all the usages of a word perfectly.



#	Keywords
0	strain coronavirus, new strain, city wuhan, novel strain, strain virus, chinese city
1	strain health, strain resources, stream, network infrastructure, international resources, likely strain
2	new strain, acute respiratory, 2019 ncov, respiratory syndrome, severe acute, identified humans
3	financial strain, feeling strain, strain coronavirus, economic strain, signs strain, strain said
4	ease strain, putting strain, strain health, reduce strain, care system, strain hospitals

Figure 2: Cluster distributions per month and top keywords for each cluster for word *strain*.

The keywords and cluster distributions for the word *strain* can be found in Figure 2. This is a polysemic word with two main senses in our corpus: as the variant of a virus or bacteria (biological term) and as "a severe or excessive demand on the strength, resources, or abilities of someone or something" (Oxford dictionary). Clusters 1, 3 and 4, which roughly match the second sense of the word (strain on healthcare systems in cluster 4, financial strain in cluster 3 and strain on resources and infrastructure in cluster 1), grow bigger across time, while clusters 0 and 2, which match the first sense of the word (e.g., new virus strain), shrink. This behavior underlines the evolution of the concerns related to the pandemic in the newspapers.

2.4 Analysis of discourse dynamics with topic modelling techniques

We continued developing methods to investigate discourse dynamics in large news collections. In Marjanen et al. (2021), applicability of two topic modelling techniques to the study of discourse change in the 19th century Finnish newspapers has been studied. The paper focused on two use cases, namely church and religion discourse, and education discourse. These two use cases demonstrate different trends in the data: decline and increase, which has been previously studied in historical literature but has not yet been approached in data-driven research. From the methodological point of view, the paper proposed several novelties: first, a combined sampling, training and inference procedure for applying topic models to large and imbalanced diachronic text collections; second, a method to quantify topic prominence for a period and thus to generalize document-wise topic assignment to a discourse level. Most importantly, the paper raises an issue of topic stretching in the dynamic topic model that limits its applicability to tracing discourse dynamics.

Latent Dirichlet Allocation (LDA) and Dynamic Topic Modelling (DTM) can fairly reliably grasp many

semi-coherent themes in past discourse and help us study the dynamics of discourses. However, both methods require a very strong interpretative element in analysing discourses. DTM is much more prone to stretch or even merge topics, which requires an interpretative assessment of whether the stretching highlights interesting continuities or if it hides discontinuities that would require attention. For LDA, stretching is not so much a problem, but often it seems interpretation is needed in seeing which topics logically relate to one another. Use cases show that LDA may provide a more reliable quantification of discourse dynamics than DTM.

Though real use cases are valuable to test practical applicability of the developed methods, they are not suitable for numerical evaluation. One of the main obstacles is the lack of training data and fundamental difficulty to annotate corpus-level phenomena. To overcome this difficulty, in Duong et al. (2021) a novel evaluation framework for discourse dynamic studies has been proposed. The idea is to exploit manually assigned article categories, available in many news corpora. Distinct periods and spikes in the data could be mimicked by sampling from a certain label according to a certain pattern, while all other categories are sampled randomly. Synthetic datasets allow for training and evaluation of models able to find a subset of documents that are related to the same theme and follow the pattern, without looking at the manually assigned labels. This allows for evaluation, comparison, and improvement of the methods, impossible on most typical use cases where ground truth is not accessible.

In Duong et al. (2021), a combination of clustering with a neural sequence-to-sequence model was proposed to extract non-stable trends and find periods of instability in the data. The best-performing method yields 78% accuracy in non-stable trend detection and 73% RandIndex for pivot time points detection. In addition it was demonstrated that a model trained on synthetic data is able to find change in real news content, without any adjustments to the data.

More detailed description of the approach and the analysis are described in our papers (Marjanen et al., 2021; Duong et al., 2021), provided in Appendices G and H.

2.5 Ideological viewpoint detection experiments on the LGBTIQ+ news corpus

We employed the methodology described in Section 2.2 also for detection of the ideological viewpoint of media. The main idea here is to compare distinct word usage distributions in media with different ideological background.

First, we collected a corpus from the Event registry (Leban et al., 2014) dataset by searching for Slovenian articles from 2014 to (including) 2020, containing any of the manually defined 125 keywords (83 unigrams and 42 bigrams) and their inflected forms connected to the subject of LGBTIQ+. The resulting corpus contains news articles on the LGBTIQ+ topic from 23 media sources. The corpus statistics are described in Table 3. Out of this corpus, we extracted a subcorpus appropriate for the viewpoint analysis. The subcorpus we used included the following online news media: Delo, Večer, Dnevnik, Nova24TV, Tednik Demokracija and PortalPolitikis. The sources were divided into two groups. The first group, namely Delo, Večer and Dnevnik represent the category of daily quality news media that are published online and in print with a long tradition in the Slovene media landscape. These three media are relatively highly trusted by readers and have the highest readership amongst Slovene dailies. The second group of news media - namely, Nova24TV, Tednik Demokracija and PortalPolitikis have been established more recently and are characterised by their financial and political connections to the Slovene right-wing/conservative political party SDS (Slovenska demokratska stranka) and the Roman Catholic Church.

The viewpoint analysis was conducted by finding words, whose usage varies the most in the two groups of media sources selected for the analysis (i.e. Delo, Dnevnik, Večer vs. Nova24TV, Tednik Demokracija and PortalPolitikis). The 10 most changed words are presented in Table 4. The word that changed the most was globok (deep), for which our system for interpretation of the change revealed that it was selected due to frequent mentions of *deep state* in the media with connections to political right. The

Table 3: LGBTIQ+ corpus statistics.

Source	Num. articles	Num. words
MMC RTV Slovenija	1790	1,555,977
Delo	1194	1,064,615
Nova24TV	844	683,336
Večer	667	552,195
24ur.com	661	313,794
Dnevnik	592	262,482
Siol.net Novice	549	460,561
Slovenske novice	501	236,516
Svet24	430	286,429
Mladina	394	275,506
Tednik Demokracija	361	350,742
Domovina	327	283,478
Primorske novice	255	183,624
Druzina.si	253	149,761
Vestnik	242	263,737
Časnik.si - Spletni magazin z mero	239	280,339
Žurnal24	172	79,953
PortalPolitikis	157	111,683
Revija Reporter	102	62,429
Gorenjski Glas	97	92,751
Onaplus	79	104,343
Športni Dnevnik Ekipa	67	33,936
Cosmopolitan Slovenija	57	71,538

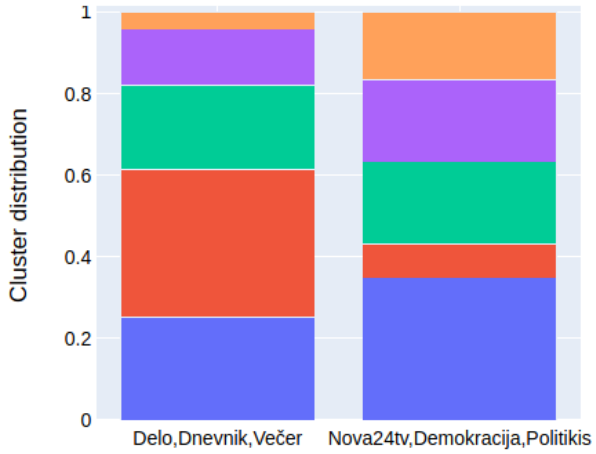
Table 4: Top 10 most changed words (and their English translations) in the LGBTIQ+ corpus according to Wasserstein distance between k-means ($k = 5$) cluster distributions in distinct chunks of the corpus.

1	globok (deep)	6	napaka (mistake)
2	roman (novel)	7	nadaljevanje (continuation)
3	video	8	lanski (last year)
4	razmerje (relationship)	9	kriza (crisis)
5	teorija (theory)	10	pogledat (look)

context of *deep state* is interesting, since it is a very frequently used interpretative frame by this group of media sources, regardless of the specific topic. Here it indicates the framing of the LGBTIQ+ questions as part of a political agenda driven by the left-wing politics.

The second word roman (novel) was selected because it appears in two contexts: as a novel and also as a constituent word in a name of the Slovenian LGBTIQ+ activist, Roman Kuhar. While the third word, *video*, is a corpus artefact that offers little insight into the attitude towards LGBTIQ+, the fourth word, *razmerje* (relationship), has a direct connection to some of the most dividing LGBTIQ+ topics, such as gay marriage, therefore for this word we provide a more detailed analysis. Figure 3 presents cluster distributions per two media groups and top 5 (translated) keywords for each cluster for word *razmerje (relationship)*. The main difference between the two distributions can be observed when it comes to mention of relationship in the context of family and marriage (see the red cluster), which present a large cluster of usages in the mainstream media but a rather small cluster in the right-wing media. On the other hand, relationship is in these media mentioned a lot more in the context of partnership, homosexuality and polygamy (see the orange cluster). The other three clusters (i.e. usages) have a rather strong presence in both media groups.

More detailed description about the approach and the analysis are described in our paper (Martinc et al., 2021), provided in Appendix B.



#	Keywords
	partnership, homosexual relationship, polygamous relationship, intimate relationship, problem
	family relationship, end work relationship, woman, field, less
	power in the relationship, relationship between people, percentage, man and woman, story
	family relationship, marriage, children, Ljubljana, coalition for children
	woman, love relationship, work relationship, film, meet

Figure 3: Cluster distributions per two media groups and top 5 translated keywords for each cluster for word *razmerje(relationship)*.

2.6 Ideological viewpoint detection experiments on the Slovenian COVID-19 news corpus

In an interdisciplinary study between NLP researchers and a sociologist, we have also investigated Slovenian media reporting around the topic of Covid-19. The paper was presented in the scope of Slovenian sociological society meeting and published in the conference proceedings (Pollak et al., 2021).

First, we collected a corpus of news article using the Event registry service (Leban et al., 2014) by selecting the following keywords: *covid, koronavirus, sars-cov-2, covid19, covid-19, korona virus, koronavirusna, koronavirusen*. We included all the articles returned by the service under the condition that the media source was registered in the Media Registry of the Slovenian Ministry of Culture. This resulted in 89.204 articles from 50 media (Corpus-50). Next, we excluded the portals specialised in sports news and local news, and kept only the media sources that contained at least 1000 topical articles. This smaller subcorpus (Corpus-14) contained the articles for the following Slovenian news portals: rtvslo.si, siol.net, delo.si, žurnal24.si, vecer.com, 24ur.com, novice.svet24.si, reporter.si, dnevnik.si, demokracija.si, nova24tv.si, politikis.si, mladina.si and necenzurirano.si.

The aim of this study was to reveal differences in news reporting between various media sources, using three different approaches. First, we analysed for which topics the reporting varies the most. Second, we analysed relatedness of different concepts with COVID-19 across different media. Finally, we analysed how word usage differs between different media and provides new insight into media reporting. We used LDA topic modelling, and methods based on contextual representations using Slovenian RoBERTa (Ulčar & Robnik-Šikonja, 2020).

For finding topical variation, we performed the following experiments. First, we applied standard LDA topic modelling (Blei et al., 2003) on the entire Corpus-50, to extract various topics in reporting COVID-19 media discourses. First, standard preprocessing techniques were used, including lemmatization, lower-casing, stop-word removal and TF-IDF weighting. Next, we applied LDA and extracted 20 topics, described by 10 words each. These topics were then manually checked, and after removing noisy topics, as well as topics which were not deemed interesting for the analysis (e.g., topics focusing exclusively on sports), we kept 12 topics that we manually named as: *closure of public life, state aid, epidemic, distance education (school), vaccination, Janša government, Croatia, testing, tracking the infection and management of epidemics, culture, measures, stock market.* The original Slovene topics and corresponding keywords can be found in Figure 4.

Next, we compute topic variation in the following way. For each word describing a specific topic, we

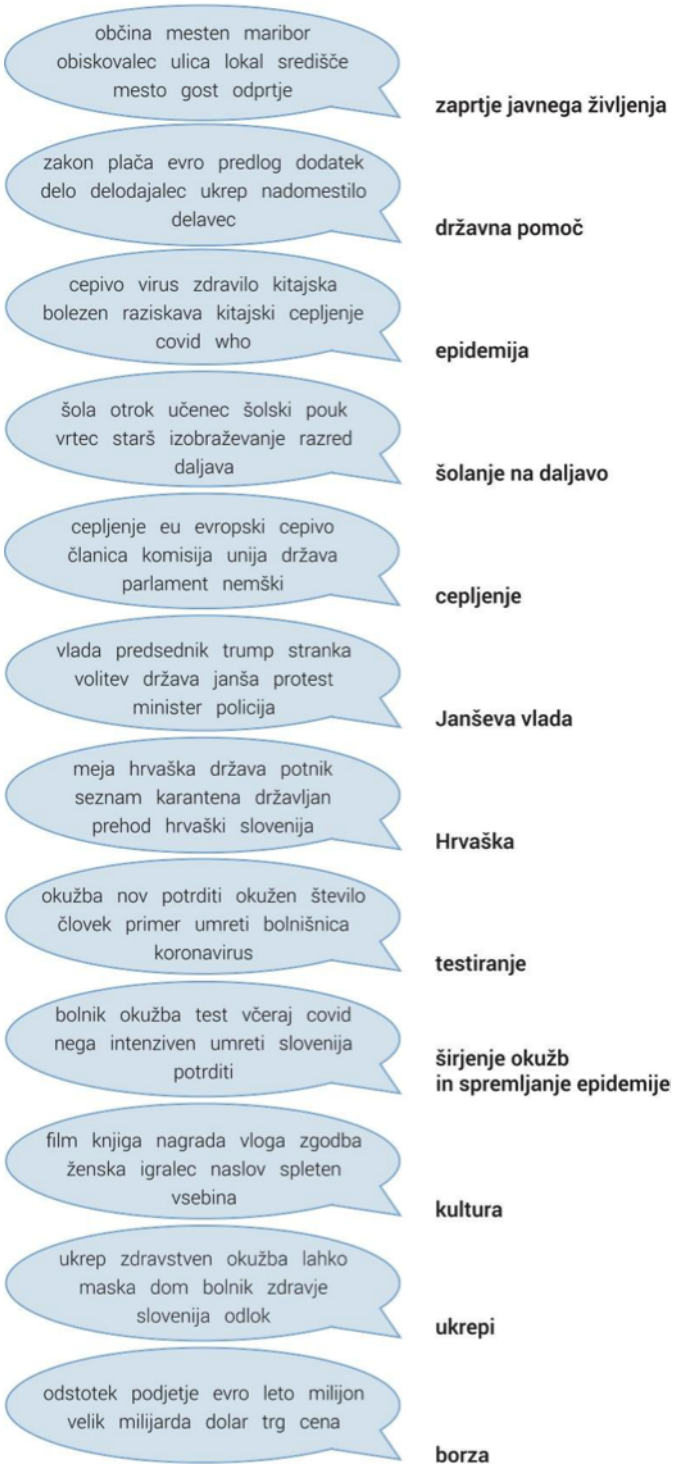


Figure 4: Topic modelling results on the Slovenian COVID-19 corpus of 50 media.

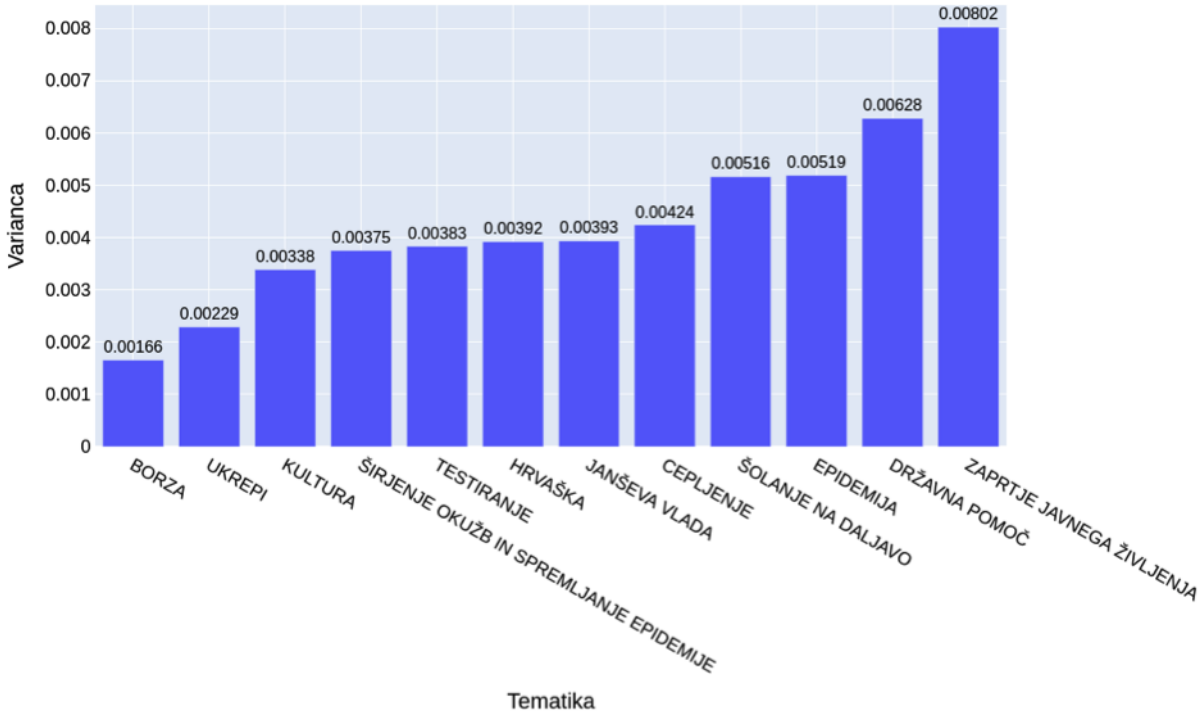


Figure 5: Variance (varianca in Slovenian) of selected topics (tematika in Slovenian) across media.

average their lemmas' contextual embeddings from Slovenian RoBERTa (Ulčar & Robnik-Šikonja, 2020) from each media source. Then, by computing the variance between the 14 selected media (Corpus-14), we can evaluate how different is the reporting on the topic for each media. The results in Figure 5 show that the topic of closure of public life and state aid have the most heterogeneous coverage based on various media sources, while the topic of stock market is the most uniformly reported on. Next, we investigated similarities and differences in media reporting by computing the relatedness between selected words and COVID-19. We selected a subset of concepts obtained by topic modelling, and added several concepts that were supposed to show distinct differences between media. The final selection included the word *vaccination (sl. cepljenje)* as it is clearly strongly related with COVID-19. Next we were interested in relatedness of *school (sl. šola), economy (sl. gospodarstvo)* and *army (sl. vojska)* to COVID-19. Finally, we investigated the relatedness of *protest (sl. protest)* and cyclist (sl. kolesar) to COVID-19. These two words were selected as the coverage is expected to be much more media-source dependant, as they relate to an anti-government movement, where protesters cycle through Slovenia's capital Ljubljana on Fridays to protest against the government.

Representations of selected words were obtained by using Slovenian RoBERTa, by averaging the words' lemmas from all of the contexts for each media source separately (we followed the approach proposed in Martinc, Novak, & Pollak (2020)). As it can be seen from Figure 6, all the news portals show the strongest correlation (in terms of cosine distance) between COVID-19 and vaccination, which is expected. One of the interesting results is that economy is more strongly related to COVID-19 than school, even though discussions about distant education were present in media reporting. The largest differences were observed for the words with stronger ideological connotations. For example, the word cyclist has stronger correlation with COVID-19 in reporting of portals nova24tv.si, demokracija.si and necenzurirano.si. This can be explained by the politicisation of the word *cyclist* in the media, which on the one hand tried to discredit the cyclist anti-government initiative (in nova24tv.si and demokracija.si), as well as in the media that were very favourable to this movement (necenzurirano.si). The word *kolesar (eng. cyclist)*) obtained with this movement a novel connotative meaning (connotation to the movement 'fighting government repression' on the one hand, and connotation to the 'destructive movement trying

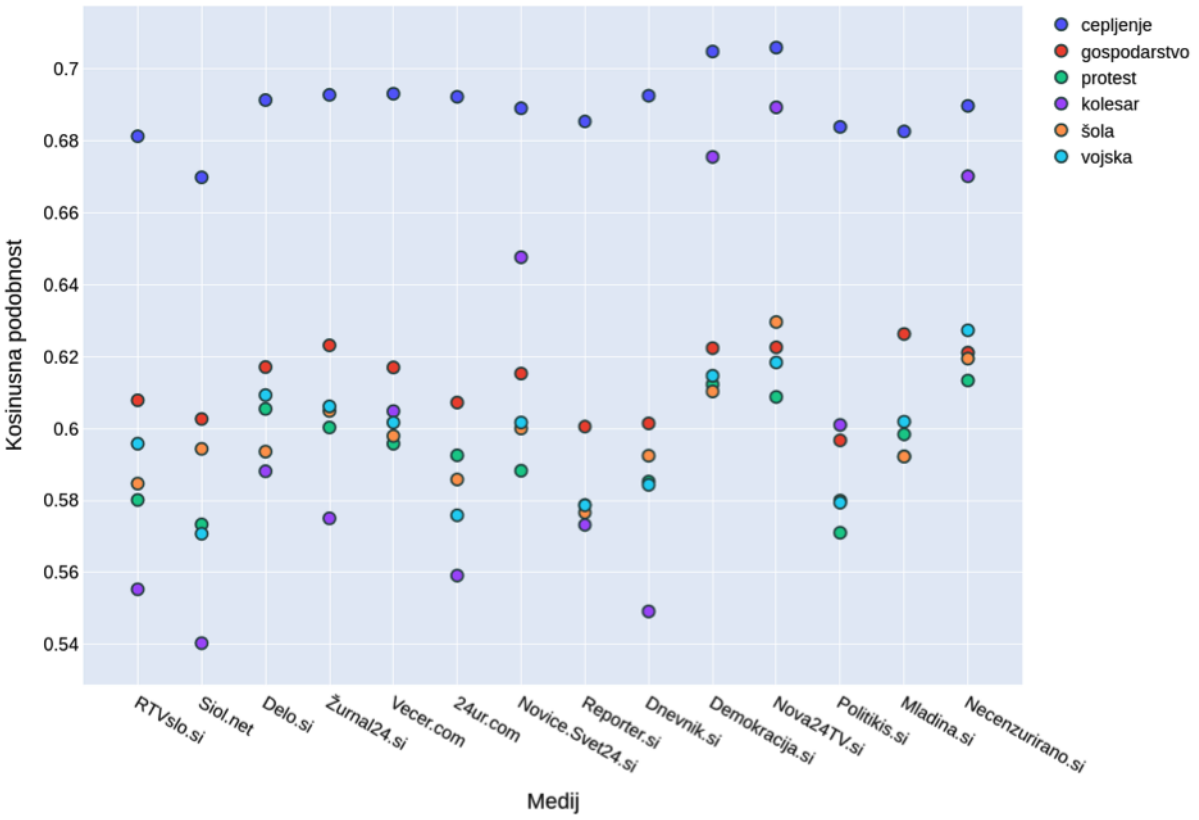


Figure 6: Relatedness of selected words to COVID-19 across different media (from Corpus-14) according to cosine distance (kosinusna podobnost in Slovenian).

to destroy the established government at whatever cost‘ on the other hand).

Last but not least, a comparative analysis via word meaning clustering was performed. We followed an approach similar to Martinc et al. (2021) and the LGBTIQ+ corpus analysis described in the previous section. Contextual representations from the corpus of 14 media sources (Corpus-14) are first grouped using k-means clustering (k is set to 5), and next, each cluster is described with keywords obtained with a simple TF-IDF based keyword extraction. The differences in the distribution of word's clusters are expected to reveal the differences in reporting across different media. Figure 7 shows cluster distributions for word *kolesar (eng. cyclist)*. Cluster 1 reflects the political meaning of word cyclist, while other clusters reflect the traditional, sports-related word meaning. It can be observed that Cluster 1 is much more represented in the news of the portals demokracija.si and nova24tv.si, followed by necenzurirano.si. On the other hand, this cluster is under-represented on the portal siol.net.

The same method also led to interesting findings about the meaning distribution of the word 'economy', where the cluster containing words related to the very disputed topic of buying of COVID-19 protective equipment by the Agency of Republic of Slovenia for Commodity Reserves, is strongly represented in the media necenzurirano.si and mladina.si, two left wing media that criticized the government. On the other hand, the same cluster is under-represented in the tabloid novice.svet24.si, which is known to be ideologically closer to the current government. The Figure is omitted from this report, but can be found in Appendix D, where also a more detailed analysis (in Slovene) is presented.

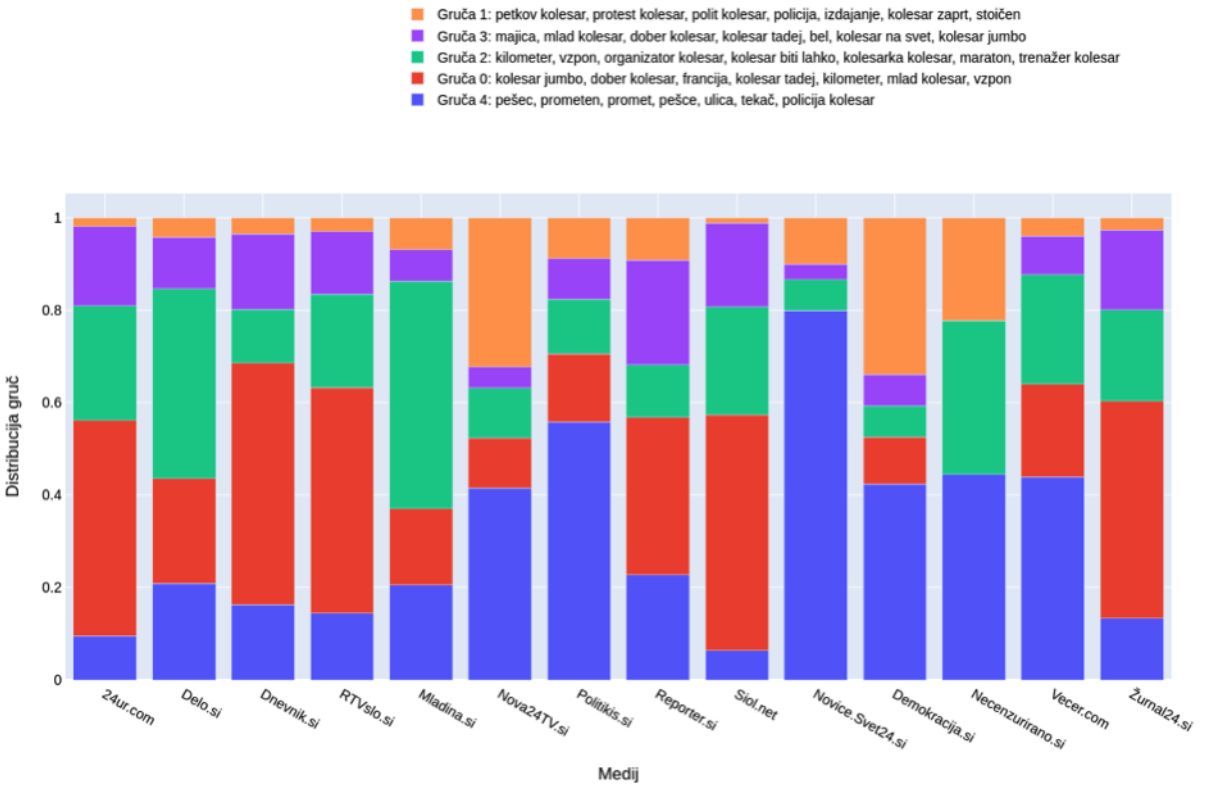


Figure 7: Cluster (Gruča in Slovenian) distributions for word kolesar (eng. cyclist) across media.

3 Sentiment analysis

In the previous deliverable for T4.3, we have first presented the Slovenian monolingual models for news sentiment analysis based on a combination of sparse and dense text representations. The experiments performed with these models have shown comparable results to the state-of-the-art models from related work. We proceeded with the experiments which explored the possibility of boosting the performance of our Slovenian monolingual sentiment classification models through the transfer learning paradigm. Using a slightly modified model architecture from the previous experiment, we first trained the models on a news categorization task before additionally training them on our target task of sentiment classification. The proposed transfer learning approach achieved comparable performance with the neural model from the previous experiment, however we did not achieve the boost in performance from the additional pre-training we initially hoped for. Finally, we conducted experiments with multilingual BERT models that could potentially be used in a cross-lingual setting in an effective manner. We have, however, identified a potential drawback of the BERT-based models for the task of sentiment news classification. As the input to the BERT-based models is limited to the 512 tokens, the standard way of preprocessing long documents is to cut the document short to a maximum length which potentially incurs the loss of valuable information. To limit the negative impact of cutting documents to a maximum length, we have experimented with several proposed long document representations and tested them in a zero shot crosslingual setting. These experiments are described in detail in deliverable D4.4.

In this deliverable we mainly focus on contributions on model adaptation using our proposed approach of sentiment-enrichment for the BERT-based models. The proposed approach was thoroughly tested in a zero-shot crosslingual setting. This work builds on the experiments on long text representations which was described in detail in deliverable D4.4.

3.1 Background and related work

Sentiment analysis is one of the most popular applications of natural language processing (NLP) and has found many areas of applications in customers' product reviews, survey textual responses, social media, etc. It analyzes users' opinions on various topics, such as politics, health, education, etc. In sentiment analysis, the goal is to analyze the author's sentiments, attitudes, emotions, and opinions (Beigi et al., 2016). Traditionally, such analysis was performed towards a specific entity that appears in the text (Mejova, 2009). A less researched, but nevertheless prominent field of research in sentiment analysis is to shift the focus from analyzing sentiment towards a specific target to analyzing the intrinsic mood of the text itself. Several works try to model feelings (positive, negative, or neutral) that readers feel while reading a certain piece of text, especially news (Bučar et al., 2018; B. Liu, 2012). In Van de Kauter et al. (2015), the authors claimed that the news production directly affects the stock market as the prevalence of positive news boosts its growth and the prevalence of negative news impedes it. In the context of news media analytics, the sentiment of news articles has been used also as an important feature in identifying fake news (Bhutani et al., 2019) and biases in the media (El Ali et al., 2018). Rambaccussing & Kwiatkowski (2020) explored the change in sentiment of news articles from major U.K. newspapers with respect to current economic conditions. Bowden et al. (2019) took a step further and tried to improve the forecasting of three economic variables, inflation, output growth, and unemployment, via sentiment modeling. They concluded that, using sentiment analysis, out of the three variables observed, the forecasting can be effectively improved for unemployment.

Recently, the use of pre-trained Transformer models has become standard practice in modeling text classification tasks. Among the first such models was the BERT (Bidirectional Encoder Representations from Transformers) model developed by Devlin et al. (2019). The initial model was however pre-trained only on English corpora and could consequently be used only for modeling textual data in the English language. A new version of the BERT model, titled multilingual BERT or mBERT, soon followed. This model was pre-trained on unlabeled data in 104 languages with the largest Wikipedias using a joint vocabulary. Several studies noted the ability of the mBERT model to work well in multilingual and cross-lingual contexts even though it was trained without an explicit cross-lingual objective and with no aligned data (Pires et al., 2019; Karthikeyan et al., 2019).

In the context of sentiment analysis of news articles, we however identified a potential drawback of the mBERT model: the input representations produced by the Transformer models may encode only a small amount of sentiment information. The pre-training objectives, namely the masked language modeling and next sentence prediction, are designed to focus on encoding general syntactic and certain semantic features of a language. The only explicit sentiment signal the models get is during the fine-tuning phase, when the models are generally trained on a much smaller amount of data.

3.2 Methodology

In the proposed approach, the aim is to induce sentiment information directly into the vectorized document representations that are produced by the multilingual BERT model. To do so, we added an intermediate training step for the mBERT model before the fine-tuning phase. The intermediate training phase consists of jointly training the model on two tasks. The first task we used was the masked language modeling task as described in the original paper by Devlin et al. (2019). We left this task unchanged in hopes that the model would better capture the syntactic patterns of our training language and domain.

For the second task, we used the sentiment classification task, which mirrors the fine-tuning task, but is trained using a different set of labeled data. With this task, we tried to additionally constrain the model to learn sentiment-related information before the actual fine-tuning phase. The task was formally modeled as a standard classification task where we tried to learn a predictor that would map the documents to a discrete number of classes:

$$\gamma : x \rightarrow C$$

For each document x_i in the training set $S = \{x_1, x_2, \ldots, x_n\}$, we produced a document representation $d \in R^{1\times t}$, where t is the dimension of the representation, by encoding the document with the mBERT model and taking the representation of the [CLS] token from the last layer. We sent this representation through a linear layer and a softmax function to map it to one of the predefined classes $C = \{y_1, y_2, \ldots, y_n\}$.

$$h = Linear(d, W) \tag{1}$$

$$\hat{y} = Softmax(h) \tag{2}$$

We calculated the loss of the sentiment classification task: $\mathcal{L}_s$ at the end using the negative log likelihood loss function

$$\mathcal{L}_s = -\log(\hat{y}_i)$$

where $\hat{y}_i$ is the probability of the correct class.

The final loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_s$$

where $\mathcal{L}_{mlm}$ represents the loss from the masked language modeling task. The model is then jointly trained on both tasks by backpropagating the final loss through the whole network.

The original mBERT model is pre-trained on another task, namely next sentence prediction, which, according to the authors, helps the model learn sentence relationships. During training, the input for this task is treated as belonging to two separate sequences and the model has to decide if the two sequences follow one another in the original text or not. This information is useful for a variety of downstream tasks such as question answering. Since in this experiment we are dealing with a classification task, where the input is treated as being a part of the same sequence, we felt the additional training using the next sentence prediction task would not add much relevant information to the model so we omitted it in the intermediate training phase.

The training of the model is divided into two phases: a sentiment enrichment intermediate training phase and a fine-tuning phase. Both phases of the training share the same training hyperparameters suggested in related work (Devlin et al., 2019). We used the Adam optimizer with the learning rate of $2E-5$ and learning rate warmup over the first 10% of the training instances. For regularization purposes, we used the weight decay set to 0.01. We reduced the batch size from 32 to 16 due to the high memory consumption during training, which was the result of a long sequence length.

We evaluated the performance of our method using Slovenian news dataset for training in a 10-fold crossvalidation setting. For the intermediate training phase, we used the Slovenian news dataset with annotations on the paragraph level. The annotations on this level of granularity were used because we wanted to perform the intermediate training phase on a different dataset than the one used for fine-tuning, but containing information relevant for the document-level sentiment classification task. Since the annotated paragraphs were part of the same documents we used for the fine-tuning step, we took measures to prevent any form of data leakage. We performed the intermediate training in each cross-validation step, but excluded the paragraphs that were part of the documents in the k-th testing fold of the fine-tuning step from the dataset. We split the remaining data into a training and development set and trained the language model for a maximum of five epochs. At the end of each epoch, we calculated the perplexity score of the model on the development set and saved the new weights only if perplexity improved in the previous epoch. If perplexity did not improve for three consecutive epochs, we stopped the training early. During the fine-tuning step, the sentiment-enriched models in each cross-validation step were trained for 3 epochs. To avoid overfitting, we split the training folds into smaller training and development sets. After each epoch, we measured the performance on the development set and saved the new model parameters only if the performance of the model on the development set increased.

3.3 Experiments

The trained models were evaluated on their respective crossvalidation fold of the Slovenian news dataset as well as on the Croatian news sentiment dataset in a zero shot setting, i.e. without any additional gradient-updating steps. The results for the intermediate training experiment (see Table 5) show that the model with the additional intermediate training step outperforms the baseline model without the intermediate training step when using the same document representation technique. The results show three points higher average performance on the Slovenian dataset and 2.68 points average improvement on the Croatian dataset in terms of the F1 score. Our model also manages to outperform the previous state-of-the-art models on the Slovenian dataset, achieving a 0.36 point increase in terms of F1 score, however this should be taken with precaution as the two evaluation settings differ. The detailed description of the method as well as the results and associated discussion can be found in Appendix C.

Table 5: Performance of the model using our intermediate sentiment classification training approach compared to the model without intermediate training. The models were tested on the Croatian dataset in a zero-shot setting. Additionally, we include the reported results from the related work using the same dataset. Best results are marked in bold.

Model	Slovenian	Croatian
	F1	**F1**
Majority classifier	22.76	25.00
Reported results from related studies		
SVM (from Bučar et al. (2018)) 5 $\times$ 10 CV	63.42 $\pm$ 1.96	/
NBM (from Bučar et al. (2018)) 5 $\times$ 10 CV	65.97 $\pm$ 1.70	/
LSTM+TF-IDF (from Pelicon (2019)) train-set split	62.5	/
Results from the current study		
No sentiment enrichment	63.34 $\pm$ 2.29	52.06 $\pm$ 2.64
With sentiment enrichment	**66.33** $\pm$ 2.60	**54.77** $\pm$ 1.39

3.3.1 Use case: Sentiment analysis on the LGBTIQ+ corpus

The model described in Section 3.2 was also applied to the LGBTIQ+ corpus described in Section 2. We focused on the differences in sentiment of reporting between well established media with long tradition of news reporting and more recently established media characterised by their financial and political connections to the Slovene conservative political party SDS. Each news article was labeled with one of the sentiment labels, namely negative, neutral or positive. This allowed us to generate a sentiment distribution of articles for each media source in the corpus.

Figure 8 presents sentiment distribution across articles for each specific news media, arranged from left to right according to the share of articles with negative sentiment. Note that all three media houses selected for the viewpoint analysis (Nova24TV, Tednik Demokracija and PortalPolitikis) because of their financial and political connections to the Slovene right-wing/conservative political party SDS produce more news articles with negative sentiment on the topic of LGBTIQ+ than the mainstream media with the long tradition (Delo, Dnevnik, Večer). The source with the most negative content about LGBTIQ+ is Revija Reporter, which is in most media analyses positioned in the right-wing ideological spectrum[14] Milosavljević (2016); Milosavljević & Biljak Gerjevič (2020). On the other side the source with the smallest share of negative news is Primorske novice, a politically independent daily regional quality news

[14] https://podcrto.si/mediji-martina-odlazka-1-del-nepregledna-mreza-radiev-tiskovin-televizije/

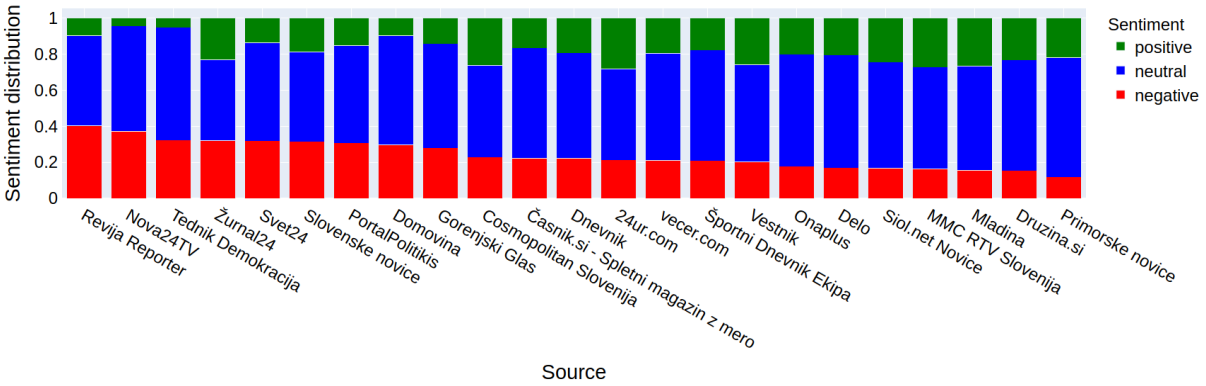


Figure 8: Sentiment distribution for each source in the LGBTIQ+ corpus.

media published online and in print with a long tradition in the regional media landscape. Nevertheless, not all conservative media are characterized by a more negative reporting about the LGBTIQ+ topic. For example, the source with the second lowest share of negative news is Druzina.si, which is strongly connected to Roman Catholic Church.

More detailed description about the approach and the analysis are described in our paper Martinc et al. (2021), provided in Appendix B.

4 Fake news identification

In deliverable D3.4, we have already proposed selected methods for fake news identification. In that deliverable we mostly focused on social media context (identification of fake news spreaders on Twitter (Koloski et al., 2020) and detection of COVID-19 related misinformation on social media (Koloski, Stepišnik-Perdih, et al., 2021)). That deliverable also included a description of our approach to the CONSTRAINT 2021 shared task (Koloski, Stepišnik-Perdih, et al., 2021). Recently, we were invited (as one of 15 out of 168 teams) to extend our work for a journal paper and the paper is currently under review (Koloski, Stepišnik-Perdih, et al., 2021). Therefore, here we propose several extension of our initial approach, such as inclusion of new datasets (including news datasets), and several novel methods.

Traditionally, to detect and remove fake news and comments media and social networks employed human curators that manually filtered the possible threats. However, this became unfeasible due to sheer scope of the problem, with millions of comments and news posts written every day, and human curators find it harder and harder to detect and filter out fake news. The spread of such news impacts many points of society, for example, it affected the elections in democratic societies (the presidential elections in USA 2016 are believed to be a major target of fake-news spread) and the well-being of humans (harmful drugs were taken as a precaution against COVID virus). In the scope of our work, we have tackled two different types of problems: the spreader detection problem and the fake-news classification problem. The first problem deals with detection of users that spread fake news, given a collection of posts written by them. The second problem deals with the classification of a single post or news article as either fake or real. In the scope of the EMBEDDIA project (and more specifically, in the scope of this deliverable), we provide several solutions to both of these problems.

4.1 Data sets

In order to evaluate our methods we use four datasets, covering different fake news problems. More specifically, 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 in this work.

COVID-19 Fake News detection data set (Patwa, Sharma, et al., 2021; Patwa, Bhardwaj, et al., 2021) 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 three different sets: train, validation and test.

Liar, Liar Pants on Fire (Wang, 2017) represents a subset of PolitiFact's collection of news that are labeled in 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 with six different degrees of fake news provided. For each news article, an additional metadata is provided consisting of: speaker, controversial statement, US party to which the subject belongs, what the text addresses, and the occupation of the subject.

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 to obtain a single document representing that author.

FNID: FakeNewsNet (Shu et al., 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, full text, speaker and the controversial statement are given.

The data splits are summarised in Table 6.

Table 6: 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 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, Stepišnik-Perdih, et al., 2021):

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, standard deviation of the word length in document. Additionally we counted the number of words beginning with upper and the number of words beginning 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. In the Koloski et al. (2020) solution to the PAN2020-Fake News profiling we applied the low dimensional space estimation technique. First we preprocessed the data by lowercasing the tweet content and removing the hashtags and punctuation. After that we removed the stopwords and obtained the final clean presentation. From the cleaned text, we generated the POS-tags using the NLTK library (Loper & Bird, 2002). We also employed word and character based n-grams. We calculated TF-IDF on all features and performed SVD (Halko et al., 2011) dimension reduction, reducing the representations to d dimensions. In the last step, we obtain the LSA representations of the tweets.

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), RoBERTa Y. Liu et al. (2019) and XLM Conneau & Lample (2019). The contextual representations were obtained via pooling-based aggregation of intermediary layers Reimers & Gurevych (2019).

4.3 Knowledge graph-based document representations

We continue the discussion by presenting the key novelty of this work: 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 (KG) 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 useful manner. To obtain the representations we use six knowledge graph embedding techniques: TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), QuatE (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 work, its entities. The considered KG embedding methods also aim to preserve relationships between entities (see the attached paper Koloski, Stepišnik-Perdih, et al. (2021) for details on each of these approaches).

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 above. For each technique we learn the space separately and later combine them in order to obtain the higher dimensional spaces useful for solving a given classification task.

For representing a given document, the proposed approach can consider the document text (we label this approach as KG) 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.) (we label this apporach as KG-ENTITY). 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).

The GraphVite library (Zhu et al., 2019) incorporates approaches that map aliases of concepts and entities into their corresponding embeddings. In the documents, 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 $\mathbf{e}^i$ 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 use the aggregation that prescribes all the concepts an equal weight and simply averages all concept embeddings:

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

4.4 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 7. Overall, 11

Table 7: 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

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 potential 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 4.2 - 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 Subsection 4.3 - 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. Here, we encounter two different scenarios for KG:

- LM+KG - we combine the induced KG representations with the methods explained in Subsection 4.3.
- LM+KG+KG-ENTITY - we combine the document representations, induced KG representations from the KG and the metadata KG representation if it is available.

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

4.5 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* or *hinge* loss function. We applied the learners on the three different representations scenarios.

Representation stacking with neural networks.

Since we have various representations both for the textual patterns and for the embeddings of 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 9. We tested three different neural networks for learning this task.

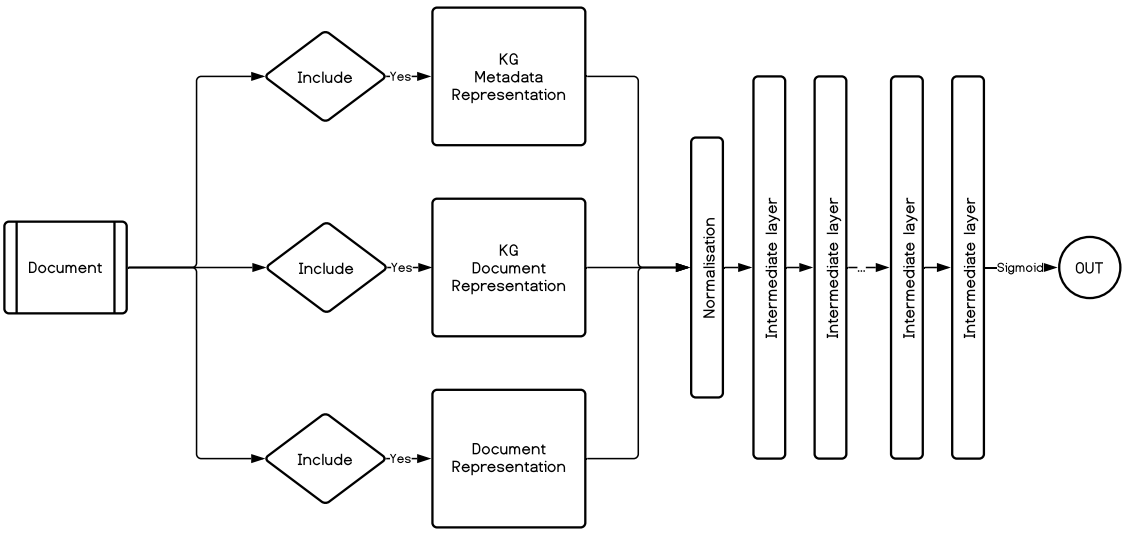


Figure 9: Neural network architecture for learning the joint intermediate representations. The *Include* decision block implies that some of the representations can be optionally excluded from the learning. The number of the intermediate layers and the dimensions are of varying sizes and are part of the model's input.

The proposed architecture consists of main two 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, Stepišnik-Perdih, 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 activation layer with the number of unique outputs.

4.6 Experiments

In this section, we evaluate and compare the quality of the representations obtained for each dataset described in Section 4.1. For each task we report four metrics: *accuracy*, *F1-score*, *precision* and *recall*.

4.6.1 Task 1: LIAR

The results are shown in Table 8. 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 the top performing 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.

Table 8: 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 continually 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**
State of the art (related work), Alhindi et al. (2018)	**0.3740**	x	x	x

4.6.2 Task 2: FakeNewsNet

The results are shown in Table 9. The Log(2) scaled neural network 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 rate 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 the recall score. The proposed methodology improves the score over the current best performing model by a margin of 3.22%.

Table 9: 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**
State of the art (related work), Bidgoly et al. (2020)	0.8558	x	x	x

4.6.3 Task 3: PAN2020

For the *PAN2020* problem, the best performing model uses the combination of the LSA and the TransE and RotatE document representations. It employs a SGD based linear model on the subsets of all of the representations learned. The deeper neural networks failed to learn the intermediate representations more successfully due to the lack of data examples (only 300 were provided). The addition of factual knowledge (embedded with the TransE and RotatE methods) to the text representation improved on the LM based representation by 10% in accuracy, and 8.59% in F1-score. The results are presented in Table 10.

Table 10: 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 StochasticGradientDescent 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**
State of the art (related work), Buda & Bolonyai (2020)	**0.7500**	x	x	x

4.6.4 Task 4: COVID-19

The model employing text based representation outperformed the model employing the KG representation according to all the metrics. However, combining the text and knowledge representations significantly improved the score. The best-performing model for this task was the SNN learned on the concatenated representation. This data set did not contain metadata information, so we ommited the KG-ENTITY evaluation. The results are shown in Table 11.

Table 11: 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**
State of the art (related work), Glazkova et al. (2020)	x	**0.9869**	x	x

Overall, 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 best-performing models were the ones that utilized the textual representations and the factual knowledge of concepts appearing in the data.

The detailed description of the method as well as the results and associated discussion can be found in Appendix E.

5 Conclusions and further work

In this deliverable we have presented our work on three topics: diachronic and ideological viewpoint detection, sentiment analysis and fake news detection. Together, these three distinct groups of methods developed in the scope of T4.3 allow for a multilingual and crosslingual analysis of news content from three distinct perspectives. The number of analysis tools developed and the number of analysis topics covered by these tools also enables a multidisciplinary news analysis serving the needs of several interested parties, e.g., the news media consumers, researchers and news media professionals.

When it comes to viewpoint detection, we proposed a scalable and interpretable method for word usage change detection, which outperforms other non-scalable contextual embeddings-based methods by a large margin. The new method also allows completely data-driven analysis of word sense dynamic in large corpora, which was impossible to conduct with unscalable methods. The applicability of the method is demonstrated on three use cases, namely on the task of diachronic word usage change detection in the COVID-19 corpus, on the task of ideological viewpoint detection in the LGBTQI+ corpus of Slovenian news, and on the task of ideological viewpoint detection in the corpus of Slovenian news about COVID-19. Additional work on the topic of diachronic viewpoint detection was conducted by exploring whether diachronic semantic change can be detected by purely grammatical features and by exploring discourse dynamic in historic news.

The results on the diachronic test corpora show that the proposed scalable and interpretable method based on contextual embeddings is in most cases outperformed by the state-of-the-art SGNS+OP+CD method. We hypothesise that this can be connected with the fact that the sentences in all but one evaluation corpus (COHA) are shuffled, meaning that BERT models cannot leverage the usual sequence of 512 tokens as a context, but are limited to the number of tokens in the sentence. We will explore this

hypothesis in the future. Despite achieving lower performance than the SGNS+OP+CD method, we nevertheless argue that our method offers a more fine-grained interpretation than methods based on non-contextual embeddings, since it accounts for the fact that words can have multiple meanings. The cluster-based technique returns a degree of change and a set of sentence clusters for each word in the corpus, roughly corresponding to word senses or particular usages. For this reason, the approach can be used for detection of new word usages (either in time or in a specific media) and for tracing how these usages disappear. Even more, word usages and their distributions over time could be linked with real-word events by labeling sentence clusters with a set of cluster-specific keywords. We will explore how to employ the proposed method for these use cases in the future.

For sentiment analysis we proposed a novel intermediate learning phase that encompasses the masked language modeling task and sentiment classification task. This phase is performed before the fine-tuning phase using a training set with separate annotations and the goal is to induce the sentiment-related information directly into the BERT representations before the fine-tuning begins on the target task data. Results show that this BERT sentiment enrichment improves the performance of the model on the Slovenian and Croatian test sets. Additionally, it slightly outperforms the current state-of-the-art on the Slovenian dataset.

In the future, we plan to further test our proposed intermediate sentiment-enrichment phase. Currently, the fine-tuning and the intermediate training phases share the dataset, but use labels on different levels of granularity: we used document-level labels for fine-tuning and paragraph-level labels for intermediate training. We would like to test how using training data from a very different training set would impact the performance of the proposed intermediate training step. We will also test the general transferability of the proposed approach to other languages and domains.

When it comes to fake news detection, we decided to tackle the fake news classification and identification of fake news spreaders with a multi-modal approach, by testing different combinations of text and knowledge graph features. We showed that combining different representations improves the overall performance of the system. We also observed during the study that knowledge graph-based representations on their own are too general for tasks where the main type of input are short texts. However, including additional statistical and contextual information about such texts has shown to improve the performance.

Otherwise, 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. For additional further work we also propose exploring attention based mechanisms to derive explanations for the specific feature significance for a classification of an instance.

6 Associated outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Availability
Montariol et al. (2021)	https://github.com/EMBEDDIA/scalable_semantic_shift	Public (MIT)
Pelicon et al. (2020)	https://github.com/EMBEDDIA/crosslingual_news_sentiment	Public (MIT)
Giulianelli et al. (2021)	https://github.com/EMBEDDIA/semchange-profiling	Public (GPL-3)
Koloski, Stepišnik-Perdih, et al. (2021)	https://github.com/EMBEDDIA/KG-informed-fake-news-classification	Public (MIT)

Parts of this work are also described in detail in the following publications, which are attached to this deliverable as appendices:

Citation	Status	Appendix
Montariol, S., Martinc, M., and Pivovarova, L. (2021). Scalable and Interpretable Semantic Change Detection. In the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, June 2021, pp. 4642–4651.	Published	Appendix A
Martinc M., Perger N., Pelicon A., Ulčar M., Vezovnik A., and Pollak S. (2021). EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+. In the Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Online, April 2021, pp. 121—126.	Published	Appendix B
Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., and Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. Applied Sciences, 10(17), 5993.	Published	Appendix C
Pollak, S., Martinc, M., Pelicon, A., Ulčar, M., Vezovnik, A. (2021). COVID-19 v slovenskih spletnih medijih: analiza s pomočjo računalniške obdelave jezika. In Pandemična družba: slovensko sociološko srečanje. Ljubljana, September 2021, pp. 260-268. (In Slovene)	Published	Appendix D
Koloski, B., Stepišnik-Perdih, T., Robnik-Šikonja, M., Pollak, S., and Škrlj, B. (2021). Knowledge Graph informed Fake News Classification via Heterogeneous Representation Ensembles. Submitted to the Neurocomputing Journal.	Submitted	Appendix E
Giulianelli, M., Kutuzov, A., and Pivovarova, L. (2021). Grammatical Profiling for Semantic Change Detection. Accepted to The SIGNLL Conference on Computational Natural Language Learning (CoNLL).	Accepted	Appendix F
Duong, Q., Pivovarova, L., and Zosa, E. (2021). Benchmarks for Unsupervised Discourse Change Detection. In Histoinformatics 2021: the 6th International Workshop on Computational History.	Published	Appendix G
Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., and Tolonen, M. (2021). Topic modelling discourse dynamics in historical newspapers. In the Post-Proceedings of the DHN2020 Conference: the 5th conference on Digital Humanities in the Nordic Countries.	Published	Appendix H

References

Alhindi, T., Petridis, S., & Muresan, S. (2018, November). Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)* (pp. 85–90). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-5513 doi: 10.18653/v1/W18-5513

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 1509–1518).

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment analysis and ontology engineering* (pp. 313–340). Springer.

Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake news detection using sentiment analysis. In *Twelfth international conference on contemporary computing (ic3)* (pp. 1–5).

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bordes, A., Usunier, N., García-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). Lake Tahoe, Nevada, United States. Retrieved from https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

Bowden, J., Kwiatkowski, A., & Rambaccussing, D. (2019). Economy through a lens: Distortions of policy coverage in uk national newspapers. *Journal of Comparative Economics*, *47*(4), 881–906.

Bučar, J., Žnidaršič, M., & Povh, J. (2018). Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, *52*(3), 895–919.

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. Retrieved from http://ceur-ws.org/Vol-2696/

Colistra, R., & Johnson, C. B. (2019). Framing the legalization of marriage for same-sex couples: An examination of news coverage surrounding the us supreme court's landmark decision. *Journal of homosexuality*.

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* (pp. 7057–7067). Vancouver, BC, Canada. Retrieved

from https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1423 doi: 10.18653/v1/N19-1423

Dubossarsky, H., Hengchen, S., Tahmasebi, N., & Schlechtweg, D. (2019, July). Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 457–470). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1044 doi: 10.18653/v1/P19-1044

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1136–1145). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/D17-1118 doi: 10.18653/v1/D17-1118

Duong, Q., Pivovarova, L., & Zosa, E. (2021). Benchmarks for unsupervised discourse change detection. In *Histoinformatics2021: the 6th international workshop on computational history.* (to appear)

El Ali, A., Stratmann, T. C., Park, S., Schöning, J., Heuten, W., & Boll, S. C. (2018). Measuring, understanding, and classifying news media sympathy on twitter after crisis events. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020, July). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3960–3973). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.365

Giulianelli, M., Kutuzov, A., & Pivovarova, L. (2021). Grammatical profiling for semantic change detection. In *The signll conference on computational natural language learning (conll).* (to appear)

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*.

Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020, July). Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 538–555). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.51

Gulordava, K., & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the gems 2011 workshop on geometrical models of natural language semantics* (pp. 67–71). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W11-2508

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, *53*(2), 217–288.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a, November). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2116–2121). Austin, Texas: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D16-1229 doi: 10.18653/v1/D16-1229

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for compu-*

tational linguistics (pp. 1489–1501). Retrieved from http://aclweb.org/anthology/P16-1141 doi: 10.18653/v1/P16-1141

Hennig, F., & Wilson, S. (2020). Diachronic embeddings for people in the news. In *Proceedings of the fourth workshop on natural language processing and computational social science* (pp. 173–183).

Kania, U. (2020). Marriage for all ('ehe fuer alle')?! a corpus-assisted discourse analysis of the marriage equality debate in germany. *Critical Discourse Studies*, *17*(2), 138–155.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual bert: An empirical study. In *International conference on learning representations.*

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* (pp. 4289–4300). Montréal, Canada. Retrieved from https://proceedings.neurips.cc/paper/2018/hash/b2ab001909a8a6f04b51920306046ce5-Abstract.html

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the acl 2014 workshop on language technologies and computational social science* (pp. 61–65). Retrieved from http://aclweb.org/anthology/W14-2517 doi: 10.3115/v1/W14-2517

Koloski, B., Pollak, S., & Skrlj, B. (2020). Multilingual detection of fake news spreaders via sparse matrix factorization. In *Clef (working notes).*

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. In *International workshop on combating on line hostile posts in regional languages during emergency situation* (pp. 177–188).

Koloski, B., Stepišnik-Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2021). Knowledge Graph informed Fake News Classification via Heterogeneous Representation Ensembles. *http://arxiv.org/abs/2110.10457*.

Kutuzov, A., & Giulianelli, M. (2020, December). UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 126–134). Barcelona (online): International Committee for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.semeval-1.14

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1384–1397). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/C18-1117

Kutuzov, A., Velldal, E., & Øvrelid, L. (2017, August). Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the events and stories in the news workshop* (pp. 31–36). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-2705 doi: 10.18653/v1/W17-2705

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: learning about world events from news. In *Proceedings of the 23rd international conference on world wide web* (pp. 107–110).

Lin, J. (2006, September). Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, *37*(1), 145–151. Retrieved from https://doi.org/10.1109/18.61115 doi: 10.1109/18.61115

Liu, B. (2012, May). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. Retrieved from http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016 doi: 10.2200/s00416ed1v01y201204hlt016

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. Retrieved from http://arxiv.org/abs/1907.11692

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 63–70). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W02-0109 doi: 10.3115/1118108.1118117

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2021). *Topic modelling discourse dynamics in historical newspapers.* (accepted)

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020a). Capturing evolution in word usage: Just add more clusters? In *Companion proceedings of the web conference 2020* (p. 343–349). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3366424.3382186 doi: 10.1145/3366424.3382186

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020b). Discovery team at semeval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 67–73).

Martinc, M., Novak, P. K., & Pollak, S. (2020). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th conference on language resources and evaluation (lrec 2020)* (p. 4811-–4819). Retrieved from http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.59.pdf

Martinc, M., Perger, N., Pelicon, A., Ulčar, M., Vezovnik, A., & Pollak, S. (2021). Embeddia hackathon report: Automatic sentiment and viewpoint analysis of slovenian news corpus on the topic of lgbtiq+. In *Proceedings of the eacl hackashop on news media content analysis and automated report generation* (pp. 121–126).

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Milosavljević, M. (2016). *Media pluralism monitor 2016 monitoring risks for media pluralism in the EU and beyond - country report: Slovenia.* Retrieved from https://cmpf.eui.eu/media-pluralism-monitor/mpm-2016-results/slovenia/

Milosavljević, M., & Biljak Gerjevič, R. (2020). *Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European Union, Albania and Turkey in the years 2018-2019 - country report: Slovenia.* European University Institute 2020. Retrieved from https://cadmus.eui.eu/bitstream/handle/1814/67818/slovenia_results_mpm_2020_cmpf.pdf?sequence=3

Montariol, S., Martinc, M., & Pivovarova, L. (2021, June). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4642–4652). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.369 doi: 10.18653/v1/2021.naacl-main.369

Paterson, L. L., & Coffey-Glover, L. (2018). Discourses of marriage in same-sex marriage debates in the uk press 2011–2014. *Journal of Language and Sexuality*, *7*(2), 175–204.

Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., PYKL, S., ... Chakraborty, T. (2021). Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Proceedings of the first workshop on combating online hostile posts in regional languages during emergency situation (CONSTRAINT).* Springer.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In *International workshop on combating online hostile posts in regional languages during emergency situation* (pp. 21–29).

Pelicon, A. (2019). Zaznavanje sentimenta v novicah z globokimi nevronskimi mrezami. *Masters thesis*.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., & Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, *10*(17), 5993.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-1202 doi: 10.18653/v1/N18-1202

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Pollak, S., Martinc, M., Pelicon, A., Ulčar, M., & Vezovnik, A. (2021, September). Covid-19 v slovenskih spletnih medijih : analiza s pomočjo računalniške obdelave jezika. In *Pandemična družba : slovensko sociološko srečanje* (pp. 260–268). Ljubljana, Slovenian: Slovensko sociološko društvo.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 231–240). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P18-1022 doi: 10.18653/v1/P18-1022

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with uk newspapers. *International Journal of Forecasting*.

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. Retrieved from CEUR-WS.org

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1410 doi: 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*. Retrieved from http://arxiv.org/abs/1910.01108

Schlechtweg, D., Hätty, A., Del Tredici, M., & Schulte im Walde, S. (2019, July). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 732–746). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1072

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020, December). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1–23). Barcelona (online): International Committee for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.semeval-1.1

Schlechtweg, D., & Schulte im Walde, S. (2020). Simulating lexical semantic change from sense-annotated data. *CoRR*, *abs/2001.03216*. Retrieved from https://arxiv.org/abs/2001.03216

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018, June). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 169–174). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-2027 doi: 10.18653/v1/N18-2027

Shi, Y., & Lei, L. (2020). The evolution of lgbt labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, *36*(4), 33–39.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019, November). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of emnlp-ijcnlp 2019* (pp. 66–76). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1007

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, *8*(3), 171–188.

Solomon, J. (2018). Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

Stewart, I., Arendt, D., Bell, E., & Volkova, S. (2017). Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. In *Eleventh international aaai conference on web and social media.*

Sun, Z., Deng, Z., Nie, J., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th international conference on learning representations, ICLR 2019.* New Orleans, LA, USA: OpenReview.net. Retrieved from https://openreview.net/forum?id=HkgEQnRqYQ

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of computational approaches to diachronic conceptual change. *CoRR*, *1811.06278*.

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, *24*(5), 649–676. doi: 10.1017/S1351324918000220

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* (Vol. 48, pp. 2071–2080). JMLR.org. Retrieved from http://proceedings.mlr.press/v48/trouillon16.html

Ulčar, M., & Robnik-Šikonja, M. (2020). Finest bert and crosloengual bert: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.

Ulčar, M., & Robnik-Šikonja, M. (2020). *Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0.* Retrieved from http://hdl.handle.net/11356/1387 (Slovenian language resource repository CLARIN.SI)

Van de Kauter, M., Breesch, D., & Hoste, V. (2015, July). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Syst. Appl.*, *42*(11), 4999–5010. Retrieved from https://doi.org/10.1016/j.eswa.2015.02.007 doi: 10.1016/j.eswa.2015.02.007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017* (pp. 5998–6008). Long Beach, CA, USA. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

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

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 422–426). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-2067 doi: 10.18653/v1/P17-2067

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. Retrieved from http://arxiv.org/abs/1412.6575

Yin, Z., Sachidananda, V., & Prabhakar, B. (2018). The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in neural information processing systems* (pp. 9412–9423).

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* (pp. 2731–2741). Vancouver, BC, Canada. Retrieved from https://proceedings.neurips.cc/paper/2019/hash/d961e9f236177d65d21100592edb0769-Abstract.html

Zheng, Y., & Chan, L. S. (2020). Framing same-sex marriage in us liberal and conservative newspapers from 2004 to 2016: Changes in issue attributes, organizing themes, and story tones. *The Social Science Journal*, 1–13.

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

Appendix A: Scalable and Interpretable Semantic Change Detection

Scalable and Interpretable Semantic Change Detection

Syrielle Montariol* LISN - CNRS, Univ. Paris-Saclay Societé Générale syrielle.montariol@limsi.fr

Matej Martinc* Jozef Stefan Institute matej.martinc@ijs.si

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

Abstract

Several cluster-based methods for semantic change detection with contextual embeddings emerged recently. They allow a fine-grained analysis of word use change by aggregating embeddings into clusters that reflect the different usages of the word. However, these methods are unscalable in terms of memory consumption and computation time. Therefore, they require a limited set of target words to be picked in advance. This drastically limits the usability of these methods in open exploratory tasks, where each word from the vocabulary can be considered as a potential target. We propose a novel scalable method for word usage-change detection that offers large gains in processing time and significant memory savings while offering the same interpretability and better performance than unscalable methods. We demonstrate the applicability of the proposed method by analysing a large corpus of news articles about COVID-19.

1 Introduction

Studying language evolution is important for many applications, since it can reflect changes in the political and social sphere. In the literature, the study of language evolution either focuses on long-term changes in the meaning of a word, or on more common short-term evolutionary phenomena, such as the word suddenly appearing in a new context, while keeping its meaning unchanged in a lexicographic sense. We refer to all types of language evolution—short- or long-term, with or without meaning change—as word usage change, a broad category that includes semantic change, but also any shifts in the context in which a word appears.

Recent studies (Giulianelli et al., 2020; Martinc et al., 2020a) show that clustering of contextual embeddings could be a proxy for word usage change: if clusters, which in theory capture distinct word usages, are distributed differently across time periods, it indicates a possible change in word's context or even loss or gain of a word sense. Thus, the cluster-based approach offers a more intuitive interpretation of word usage change than alternative methods, which look at the neighborhood of a word in each time period to interpret the change (Gonen et al., 2020; Martinc et al., 2020b) and ignore the fact that a word can have more than one meaning. The main limitation of the cluster-based methods is the scalability in terms of memory consumption and time: clustering is applied to each word in the corpus separately and all occurrences of a word need to be aggregated into clusters. For large corpora with large vocabularies, where some words can appear millions of times, the use of these methods is severely limited.

To avoid the scalability issue, cluster-based methods are generally applied to a small set of less than a hundred manually pre-selected words (Giulianelli et al., 2020; Martinc et al., 2020a). This drastically limits the application of the methods in scenarios such as identification of the most changed words in a large corpus or measuring of usage change of extremely frequent words, since clustering of all of word's contextual embeddings requires large computational resources. One way to solve the scalability problem using contextual embeddings is to *average* a set of contextual representations for each word into a single static representation (Martinc et al., 2020b). Averaging, while scalable, loses a lot on the interpretability aspect, since word usages are merged into a single representation.

The method we propose in this paper tackles scalability and interpretability at the same time. The main contributions of the paper are the following:

- A *scalable* method for contextual embeddings clustering that generates interpretable representations and outperforms other cluster-based methods.
- A method of measuring word usage change between periods with the *Wasserstein distance*. As far as we are aware, this is the first paper leverag-

* These authors contributed equally.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4642–4652
June 6–11, 2021.

ing optimal transport for lexical semantic change detection.

- A *cluster filtering* step, which balances the deficiencies of clustering algorithms and consistently improves performance.
- An *interpretation pipeline* that automatically labels word senses, allowing a domain expert to find the most changing concepts and to understand *how* those changes happened.

The practical abilities of our method are demonstrated on a large corpus of news articles related to COVID-19, the Aylien Coronavirus News Dataset[1]. We compute the degree of usage change of almost 8,000 words, i.e., all words that appear more than 50 times in every time slice of the corpus, in the collection of about half a million articles in order to find the most changing words and interpret their drift[2].

2 Related Work

Diachronic word embedding models have undergone a surge of interest in the last two years with the successive publications of three articles dedicated to a literature review of the domain (Kutuzov et al., 2018; Tahmasebi et al., 2018; Tang, 2018). Most approaches build static embedding models for each time slice of the corpus and then make these representations comparable by either employing *incremental updating* (Kim et al., 2014) or *vector space alignment* (Hamilton et al., 2016b). The alignment method has proved superior on a set of synthetic semantic drifts (Shoemark et al., 2019) and has been extensively used (Hamilton et al., 2016b; Dubossarsky et al., 2017) and improved (Dubossarsky et al., 2019) in the literature. The recent SemEval Task on Unsupervised lexical semantic change detection has shown that this method is most stable and yields the best averaged performance across four SemEval corpora (Schlechtweg et al., 2020).

Yet another approach (Hamilton et al., 2016a; Yin et al., 2018) is based on comparison of neighbors of a target word in different time periods. This approach has been recently used to tackle the scalability problem (Gonen et al., 2020).

In all these methods, each word has only one representation within a time slice, which limits the sensitivity and interpretability of these techniques.

The recent rise of contextual embeddings such as BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018) introduced significant changes to word representations. Contextual embeddings can be used for usage change detection by aggregating the information from the set of token embeddings. This can be done either through averaging of all vectors within a time slice and then computing averaged vector similarity (Martinc et al., 2020b), by computing a pairwise distance between vectors from different time slices (Kutuzov and Giulianelli, 2020), or by clustering all token representations to approximate its set of senses (Giulianelli et al., 2020). The analysis in this paper derives from this last set of methods, which demonstrate a higher performance than static embeddings methods at least on some datasets (Martinc et al., 2020a).

Automatic semantic shift detection has been used for text stream monitoring tasks, such as event detection (Kutuzov et al., 2017) viewpoint analysis (Azarbonyad et al., 2017) or monitoring of rapid discourse changes during crisis events (Stewart et al., 2017). None of these applications use clustering techniques and, as far as we are aware, only Martinc et al. (2020b) uses contextual embeddings for news stream analysis. In this paper we demonstrate the large potential of contextual embeddings for the *interpretable* tracking of short-term changes in word usage, which has a practical application for crisis-related news monitoring.

3 Scalability and Interpretability Limitations of Previous Methods

The main motivation for this research are the scalability or interpretability issues of previous methods for word usage change detection. The ones using contextual embeddings are either interpretable but unscalable (Giulianelli et al., 2020; Martinc et al., 2020a) or scalable but uninterpretable (Martinc et al., 2020b). The scalability issues of interpretable methods can be divided into two problems.

Memory consumption: Giulianelli et al. (2020) and Martinc et al. (2020a) apply clustering on all embeddings of each target word. This procedure becomes unfeasible for large sets of target words or if the embeddings need to be generated on a large corpus, since too many embeddings need to be saved into memory for further processing. To give an example, single-precision floating-point in Python requires 4 bytes of memory. Each contextual embedding contains 768 floats (Devlin et al.,

[1] https://blog.aylien.com/free-coronavirus-news-dataset/

[2] The code can be found at https://github.com/matejMartinc/scalable_semantic_shift

2019), leading each embedding to occupy 3072 bytes[3]. To use the previous methods on the Aylien Coronavirus News Dataset, which contains 250M tokens, about 768 Gb RAM would be necessary to store the embeddings for the entire corpus. If we limit our vocabulary to the 7,651 words that appear at least 50 times in every time slice and remove the stopwords (as we do in this work), we still need to generate contextual embeddings for 120M tokens, which is about 369 Gb of RAM.

Complexity of clustering algorithms: For the complexity analyses, we denote by d the dimension of the embedding, k is the number of clusters and n is the number of contextual embeddings, i.e., the number of word occurrences in the corpus. The time complexity of the affinity propagation algorithm (the best performing algorithm according to Martinc et al. (2020a)) is $O(n^2td)$, with t being the predefined maximum number of iterations of the data point message exchange. The time complexity of the simpler k-means algorithm[4] can be stated as $O(tknd)$, where t is the number of iterations of Lloyd's algorithm (Lloyd, 1982). As an example, consider the word *coronavirus*, which appears in the Aylien corpus about 1,2M times. For k-means with $k = 5$ and a maximal number of iterations set to 300 (the Scikit library default), about $300 * 5 * 1,300,000 * 768 \approx 1.5 \times 10^{12}$ operations are conducted for the clustering. With affinity propagation with the maximum number of iterations set to 200 (the default), clustering of the word *coronavirus* would require $1,300,000^2 * 200 * 768 \approx 2.6 \times 10^{17}$ operations, which is impossible to conduct in a reasonable amount of time on a high end desktop computer.

Contextual Embeddings Method with Interpretability Limitations: The averaging approach (Martinc et al., 2020b) eliminates the scalability problems: token embeddings for each word are not collected in a list but summed together in an element-wise fashion, which means that only 768 floats need to be saved for each word in the vocabulary. The averaged word representation is obtained for each time slice by dividing the sum by the word count. A single embedding per word is saved, leading to only 23.5 Mb of RAM required to store the embeddings for 7,651 words. These representations loose on the interpretability aspect, since all word usages are merged into a single averaged representation. It makes the method inappropriate for some tasks such as automatic labelling of word senses, and in some cases affects the overall performance of the method (Martinc et al., 2020a).

4 Methodology

Our word usage change detection pipeline follows the procedure proposed in the previous work (Martinc et al., 2020a; Giulianelli et al., 2020): for each word, we generate a set of contextual embeddings using BERT (Devlin et al., 2019). These representations are clustered using k-means or affinity propagation and the derived cluster distributions are compared across time slices by either using Jensen-Shannon divergence (JSD) (Lin, 2006) or the Wasserstein distance (WD) (Solomon, 2018). Finally, words are ranked according to the distance measure, assuming that the ranking resembles a relative degree of usage shift.

The primary contributions of this work lay in the embedding generation step, which improves the scalability of the method, and in leveraging WD to compute the distance between clusters. We also propose post-processing steps, which domain experts could use for the interpretation of results. We now describe the pipeline in more details.

4.1 Embeddings Generation

We use a pre-trained BERT model for each language of the evaluation corpora[5]. All models have 12 attention layers and a hidden layer of size 768. We fine-tune them for domain adaptation on each corpus as a masked language model for 5 epochs. Then, we extract token embeddings from the fine-tuned models. Each corpus is split into time slices. The models are fed 256 tokens long sequences in batches of 16 sequences at once. We generate sequence embeddings by summing the last four encoder output layers of BERT, following Devlin et al. (2019). Next, we split each sequence into 256 subparts to obtain a separate contextual embedding of size 768 for each token. Since one token does not necessarily correspond to one word due to byte-

[3]If we ignore the additional memory of a Python container—e.g., a Numpy list or a Pytorch tensor—required for storing this data.

[4]Here we are referring to the Scikit implementation of the algorithm employed in this work: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[5]For German: bert-base-german-cased (https://deepset.ai/german-bert, for English: bert-base-uncased model, for Latin: bert-base-multilingual-uncased model from the `huggingface` library, for Swedish: bert-base-swedish-uncased (https://github.com/af-ai-center/SweBERT).

pair tokenization, we average embeddings for each byte-pair token constituting a word to obtain embeddings for each occurrence of a word.

Next, after obtaining a contextual embedding vector for each target word in a specific sequence, we decide whether this vector should be *saved* to the list or *merged* with one of the previously obtained vectors for the same word in the same time slice. To improve the scalability, we limit the number of contextual embeddings that are kept in the memory for a given word and time slice to a predefined threshold. The threshold of 200 was chosen empirically from a set of threshold candidates (20, 50, 100, 200, 500) and offers a reasonable compromise between scalability and performance. The new vector is merged if it is too similar—i.e., a duplicate or a near-duplicate—to one of the saved vectors or if the list already contains a predefined maximum number of vectors (200 in our case).

More formally, we add the new embedding e_{new} to the list of word embeddings $L = \{e_i, ..., e_n\}$ if:

$$|L| < 200 \quad \& \quad \forall e_i \in L\colon s(e_{\text{new}}, e_i) < 1 - \varepsilon$$

where s is the cosine similarity and ε is a threshold set to 0.01.

If $|L| \geq 200$ or if any vector in the list L is a near duplicate to e_{new}, we find a vector e_m in the list which is the closest to e_{new} in terms of cosine similarity:

$$e_m = \arg\max_{e_i \in L} s(e_i, e_{\text{new}})$$

This element e_m is then modified by summing it with e_{new}:

$$e_m \leftarrow e_m + e_{\text{new}}$$

The number of summed-up elements for each of the 200 groups in the list is stored besides their summed-up representations. Once the model has been fed with all the sequences in the time slice, the final summed-up vector is divided by this number to obtain an averaged embedding.

By having only 200 merged word embeddings per word per time slice, and by limiting the vocabulary of the corpus to 7,651 target words, we require up to 4.7 Gb of space for each time slice, no matter the size of the corpus. While this is still 200 times more space than if the averaging method was used (Martinc et al., 2020b), the conducted experiments show that the proposed method nevertheless keeps the bulk of the interpretability of the less scalable method proposed by Giulianelli et al. (2020), and offers competitive performance on several corpora.

4.2 Clustering

After collecting 200 vectors for each word in each time slice, we conduct clustering on these lists to extract the usage distribution of the word at each period. Clustering for a given word is performed on the set of all vectors from all time slices jointly.

We use two clustering methods previously applied for this task, namely k-means used in Giulianelli et al. (2020) and affinity propagation in Martinc et al. (2020a). The main strength of affinity propagation is that the number of clusters is not defined in advance but inferred during training. The clustering is usually skewed: a limited number of large clusters is accompanied with many clusters consisting of only a couple of instances. Thus, affinity propagation allows to pick out the core senses of a word. K-means tends to produce more even clusters. Appearance of small clusters that contain only few instances and do not represent a specific sense or usage of the word is nevertheless relatively common, since BERT is sensitive to syntax and pragmatics, which are not necessarily relevant for usage change detection. Another limitation of the k-means algorithm is that the number of clusters needs to be set in advance. This means that if the number of actual word usages is smaller than a predefined number of clusters, k-means will generate more than one cluster for each word usage.

To compensate for these deficiencies, we propose an additional *filtering and merging* step. A cluster is considered to be a legitimate representation of a usage of the word, if it contains at least 10 instances[6]. We compute the average embedding inside each cluster, and measure the cosine distance (1 - cosine similarity) between the average embeddings in each pair of legitimate clusters for a given word. If the distance between two clusters is smaller than a threshold, the clusters are merged. The threshold is defined as $avg_{cd} - 2 * std_{cd}$, where avg_{cd} is the average pairwise cosine distance between all legitimate clusters and std_{cd} is the standard deviation of that distance. This merging procedure is applied recursively until the minimum distance between the two closest clusters is larger than the threshold. After that, the merging proce-

[6]The threshold of 10 was derived from the procedure for manual labelling employed in the SemEval Task (Schlechtweg et al., 2020), where a constraint was enforced that the specific sense is attested at least 5 times in a specific time period in order to contribute word senses. We set the overall threshold of 10, which roughly translates to 5 per time period, since all of our test corpora (besides Aylien) contain two time periods.

dure is applied to illegitimate clusters (that contain less than 10 instances), using the same threshold. Illegitimate clusters could be added into one of the legitimate clusters or merged together to form a legitimate cluster with more than 10 instances. If there is no cluster that is close enough to be merged with, the illegitimate cluster is removed.

4.3 Change Detection and Interpretation

After the clustering procedure described above, for each word in each time slice, we extract its cluster distribution and normalise it by the word frequency in the time slice. Then target words are *ranked* according to the usage divergence between successive time slices, measured with the JSD or the WD[7]. If a ground-truth ranking exists, the method can be evaluated using the Spearman Rank Correlation to compare the true and the outputted ranking. In the exploratory scenario, the ranking is used to detect the most changing words and then investigate the most unevenly distributed clusters over time for the interpretation of the change.

JSD has been used for semantic shift detection in several recent papers, e.g. (Martinc et al., 2020a; Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020). Since this is the first paper applying WD for this purpose, we describe it in more details.

The motivation for using the WD (Solomon, 2018) is to take into account the position of the clusters in the semantic space when comparing them. The JSD leverages semantic information encoded in the embeddings indirectly, distilled into two time-specific cluster distributions that JSD receives as an input. In addition to cluster distributions, WD accesses characteristics of the semantic space explicitly, through a matrix of cluster averages (obtained by averaging embeddings in each cluster) of size $T \times k \times 768$, where k is a number of clusters, T is a number of time slices and 768 is the embedding dimension.

This setup is a classical problem that can be solved using optimal transport (Peyré et al., 2019). We denote with μ_1 and μ_2 the sets of k average embedding points in the two vector spaces, and with c_1 and c_2 the associated clusters distributions. Thus, c_1 and c_2 are histograms on the simplex (positive and sum to 1) that represent the weights of each embedding in the source (μ_1) and target (μ_2) distributions. The task is to quantify the effort of moving one unit of mass from μ_1 to μ_2 using a chosen cost function, in our case the cosine distance. It is solved by looking for the transport plan γ, which is the minimal effort required to reconfigure c_1's mass distribution into that of c_2. The WD is the sum of all travels that have to be made to solve the problem:

$$\mathrm{WD}(c_1, c_2) = \min_{\gamma} \sum_{i,j} \gamma_{i,j} M_{i,j}$$
$$\text{with } \gamma 1 = c_1;\ \gamma^\intercal 1 = c_2;\ \gamma \geq 0$$

Where $M \in \mathbb{R}^{+}_{m \times n}$ is the cost matrix defining the cost to move mass from μ_1 to μ_2. We use the cosine similarity s, with $M = 1 - s(\mu_1, \mu_2)$.

Interpretation. Once the most changing words are detected, the next step is to understand *how* they change between two time slices by interpreting their clusters of usages.

Cluster distributions can be used directly to identify the clusters that are unevenly distributed across a time dimension. However, a cluster itself may consist of several hundreds or thousands of word usages, i.e. sentences. Interpreting the underlying sense behind each cluster by manually looking at the sentences is time-consuming. To reduce human work, we extract the most discriminating words and bigrams for each cluster: by considering a cluster as a single document and all clusters as a corpus, we compute the term frequency - inverse document frequency (tf-idf) score of each word and bigram in each cluster. The stopwords and the words appearing in more than 80% of the clusters are excluded to ensure that the selected keywords are the most discriminant. Thus, a ranked list of keywords for each cluster is obtained and top-ranked keywords are used for the interpretation of the cluster.

5 Evaluation

We use six existing manually annotated datasets for evaluation. The first dataset, proposed by Gulordava and Baroni (2011), consists of 100 English words labelled by five annotators according to the level of semantic change between the 1960s and 1990s[8]. To build the dataset, the annotators evaluated semantic change using their intuition, without looking at the context. This procedure is problematic since an annotator may forget or not be aware of a particular sense of the word.

[7] Using the POT package https://pythonot.github.io/.

[8] In order to make the proposed approach comparable to previous work, we remove four words that do not appear in the BERT vocabulary from the evaluation dataset, same as in Martinc et al. (2020a).

	COHA	SE English	SE Latin	SE German	SE Swedish	DURel	Avg. all
METHODS NOT USING CLUSTERING							
SGNS + OP + CD	0.347	0.321	0.372	**0.712**	**0.631**	**0.814**	**0.533**
Nearest Neighbors	0.310	0.150	0.273	0.627	0.404	0.590	0.392
Averaging	0.349	0.315	**0.496**	0.565	0.212	0.656	0.432
NON-SCALABLE CLUSTERING METHODS							
k-means 5 JSD	0.508	0.189	0.324	0.528	0.238	0.560	0.391
aff-prop JSD	**0.510**	0.313	0.467	0.436	-0.026	0.542	0.374
INTERPRETABLE SCALABLE METHODS							
Without filtering or merging of clusters							
k-means 5 JSD	0.430	0.316	0.358	0.508	0.073	0.658	0.390
aff-prop JSD	0.394	0.371	0.346	0.498	0.012	0.512	0.355
k-means 5 WD	0.372	0.360	0.450	0.514	0.316	0.607	0.437
aff-prop WD	0.369	**0.456**	0.397	0.421	0.264	0.484	0.399
With filtering and merging of clusters							
k-means 5 JSD	0.448	0.318	0.374	0.519	0.073	0.649	0.397
aff-prop JSD	0.403	0.348	0.408	0.583	0.018	0.712	0.412
k-means 5 WD	0.382	0.375	0.466	0.520	0.332	0.628	0.451
aff-prop WD	0.352	0.437	0.488	0.561	0.321	0.686	0.474

Table 1: Spearman Rank Correlation between system output rankings and ground truth rankings for various datasets. "SE" stands for SemEval.

The organizers of the recent SemEval-2020 Task 1— Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020)—employed another approach: the annotators had to decide whether a pair of sentences from different time periods convey the same meaning of the word (Schlechtweg and Schulte im Walde, 2020). For each of the four languages—German, English, Latin and Swedish—senses were manually annotated by labeling word senses in a pair of sentences drawn from different time periods. All SemEval-2020 Task 1 corpora contain only two periods and the sentences are shuffled and lemmatized. The lexical semantic change score is defined as the difference between word sense frequency distributions in the two time periods and measured by the Jensen-Shannon Distance (Lin, 2006).

The DURel dataset (Schlechtweg et al., 2018) is composed of 22 German words, ranked by semantic change by five annotators between two time periods, 1750–1799 and 1850–1899. Similarly to SemEval, the ranking was build by evaluating the relatedness of pairs of sentences from two periods.

In order to conduct usage change detection on the target words proposed by Gulordava and Baroni (2011), we fine-tune the English BERT-base-uncased model and generate contextual embeddings on the Corpus of Historical American English (COHA)[9]. We only use data from the 1960s to the 1990s (1960s has around 2.8M and 1990s 3.3M words), to match the manually annotated data. For the SemEval Task 1 evaluation set, we fine-tune the BERT models and generate contextual embeddings on the four corpora provided by the organizers of the task, English (about 13.4M words), German (142M words), Swedish (182M words) and Latin (11.2M words). Finally, we fine-tune BERT and generate embeddings on the German DTA corpus (1750–1799 period has about 25M and 1850–1899 has 38M tokens)[10].

The results are shown in Table 1. We compare our scalable approach with the *non-scalable clustering* methods used by Giulianelli et al. (2020) and Martinc et al. (2020a). *Averaging* (Martinc et al., 2020b) is the less interpretable method described in Section 3. *SGNS + OP + CD* (Schlechtweg et al., 2019) refers to the state-of-the-art semantic change detection method employing non-contextual word embeddings: the Skip-Gram with Negative Sampling (SGNS) model is trained on two periods independently and aligned using Orthogonal Procrustes (OP). Cosine Distance (CD) is used to compute the semantic change. The *Nearest Neighbors* method (Gonen et al., 2020) also uses SGNS embeddings.

[9] https://www.english-corpora.org/coha/

[10] https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/durel/

For each period, a word is represented by its top nearest neighbors (NN) according to CD. Semantic change is measured as the size of the intersection between the NN lists of two periods.

On average, the proposed scalable clustering with filtering and merging of clusters leads to a higher correlation with gold standard than the standard non-scalable clustering methods: the best method (aff-prop WD) achieving a Spearman correlation with the gold standard of 0.474 compared to the best non-scalable k-means 5 JSD achieving the Spearman correlation of 0.391. The method also outperforms averaging and NN, though it is outperformed by a large margin by the SGNS+OP+CD, achieving the score of 0.533.

The best performing clustering algorithm differs for different datasets. On average, affinity propagation only outperforms k-means when filtering and merging of clusters is employed. The effect of the filtering on k-means is positive on average but the difference is thin, as the number of clusters is low.

WD leads to better results than JSD on most of the corpora where averaging outperforms clustering, the only exception is DURel. An extreme example is the Swedish SemEval dataset, where the clustering with JSD performs particularly poorly: using the WD, which takes into account the average embeddings on top of cluster distributions, greatly increases the correlation with the gold standard. On the contrary, on COHA where averaging performs poorly in comparison to clustering, WD is under-performing.

6 Use Case: Aylien COVID-19 Corpus

The combination of scalable clustering with the interpretation pipeline opens new opportunities for diachronic corpus exploration. In this section, we demonstrate how it could be used to analyze the Aylien Coronavirus News Dataset. The corpus contains about 500k news articles related to COVID-19 from January to April 2020[11], unevenly distributed over the months (160M words in March, 41M in February, 35M in April and 10M in January). We split the corpus into monthly chunks and apply our scalable word usage change detection method.

6.1 Identification of the Top Drifting Words

The scalable method allows to perform embeddings extraction and clustering for all words in the corpus.

[11]We used an older version of the corpus. Currently the data from May are also available.

1	diamond	6	tag
2	king	7	paramount
3	ash	8	lynch
4	palm	9	developers
5	fund	10	morris

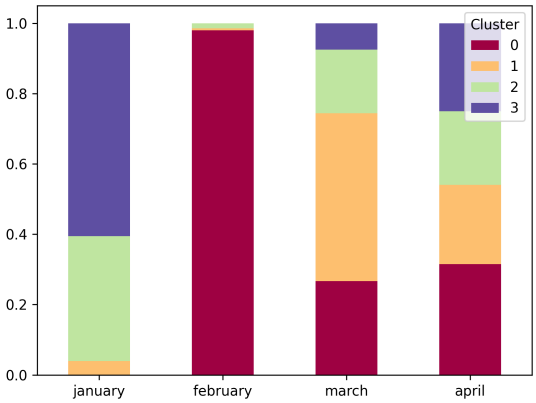
Table 2: Top 10 most changed words in the corpus according to a monthly-averaged WD of k-means ($k = 5$) cluster distributions.

We extract the top words with the highest average WD between the successive months to conduct a deeper analysis. We exclude words that appear less than 50 times in each month to avoid spurious drifts due to words having too few occurrences in a time slice. However, some drifts due to corpus artefacts remain, in particular dates such as *'2019-20'*. Thus, words containing numbers and one-letter words are also removed.

In Table 2 we present the top 10 most drifting words extracted using k-means with k=5 and ranked according to the average WD across the four months[12]. Among them, the word *diamond* is related to the cruise ship "Diamond Princess", which suffered from an outbreak of COVID-19 and was quarantined for several weeks. The word *king*, which is the second most changing word, is related to the King county, Washington, where the first confirmed COVID-19 related death in the USA appeared, and to the Netflix show "Tiger King", which was released in March. Thus, the primary context for this word changed several times, which is reflected in our results. Other words are mostly constituent words in named entities, related e.g., to an American Society of Hematology (ASH) Research Collaborative's Data Hub, which is capturing data on subjects tested positive for COVID-19.

The results suggest that the model does what it is meant to do: for most words in the list it is possible to find an explanation why its usage changed during the beginning of 2020. The list contains many proper names or proper name constituents, which could be either desirable or undesirable property, depending on research goals. Some work focuses specifically on proper names (Hennig and Wilson, 2020), since they could be a good proxy to shifts in socio-political situations. On the other hand, if

[12]This is a rather arbitrary procedure: one can imagine that a domain expert would prefer a different frequency threshold or focus more on a given month. The most time-consuming part is embedding extraction. Once this is done, clustering and keyword extraction can be done as many times as necessary.



#	Keywords
0	diamond princess, cruise ship, princess cruise, japanese, tested positive, confirm, ship diamond
1	neil diamond, comic, sweet caroline, trump, song, diamond said, comic book,
2	diamond hill, hill capital, diamond jubilee, diamond mountain, league postponed, portfolio, athletics
3	diamond industry, black diamond, jewellery, hong kong, diamond ring, surat diamond, india

Figure 1: Cluster distributions per month and top keywords for each cluster for word *diamond*.

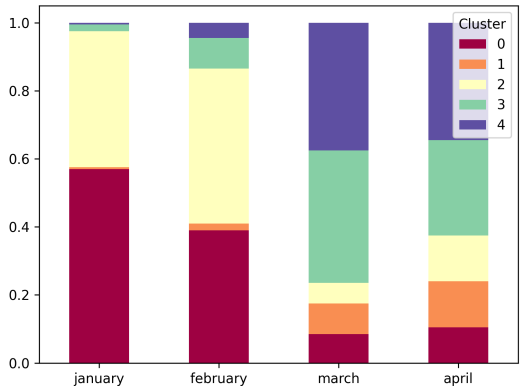
the focus of the study are shifts in more abstract concepts, then proper names could be filtered out before the embedding generation stage by employing named entity recognition tools.

6.2 Interpretation of the Usage Change

The interpretation pipeline, described in Section 4.3, is illustrated in figures 1 and 2. We focus on two words, *diamond* and *strain*, to show the various phenomena that can be detected. *Diamond* is the top drifting word in the entire vocabulary (see Table 2); it can be both a common noun and an entity, inducing usage drift when the entity appears in the newspapers after events with high media coverage. *Strain* is the 38th word with the highest drift overall, and the 15th highest between February and March 2020. It has several different senses whose usage vary across time following the events in the news. We cluster their vector representations from the Aylien corpus using k-means with $k = 5$ and apply the cluster filtering and merging step. Then, using tf-idf on unigrams and bigrams, we extract a set of keywords for each cluster to interpret the variations of their distribution.

The keywords and cluster distributions for the word *diamond* can be found in Figure 1. One of the clusters was removed at the filtering step, as it had less than 10 embeddings inside, and no other cluster was close enough. A clear temporal tendency is visible from the cluster distribution in Figure 1: a new major usage appears in February, corresponding to the event of the quarantined cruise ship (Cluster 0); this association is revealed by the keywords for this cluster. Moreover, the WD between January and February, when the outbreak happened, is 0.337; it is also very high between February and March (0.342). It reflects the large gap between the cluster distributions, first with the appearance of Cluster 0 in February that made the other usages of the word *diamond* in the media almost disappear, and then the reappearance of other usages in March, when the situation around the cruise ship gradually normalized. Cluster 1, that appears in March, is related to Neil Diamond's coronavirus parody of the song "Sweet Caroline" which was shared mid-March on the social media platforms and received a lot of attention in the US. Cluster 3 is related to the diamond industry; it is much less discussed as soon as the pandemic breaks out in February. Finally, Cluster 2 deals with several topics: Diamond Hill Capital, a US investment company, and the Wanda Diamond League, an international track and field athletic competition which saw most of its meetings postponed because of the pandemic. This last cluster shows the limitations of our clustering: it is complex to identify and differentiate all the usages of a word perfectly.

The keywords and cluster distributions for the word *strain* can be found in Figure 2. This is a polysemic word with two main senses in our corpus: as the variant of a virus or bacteria (biological term) and as "a severe or excessive demand on the strength, resources, or abilities of someone or something" (Oxford dictionary). Clusters 1, 3 and 4, which roughly match the second sense of the word (strain on healthcare systems in cluster 4, financial strain in cluster 3 and strain on resources and infrastructure in cluster 1), grow bigger across time, while clusters 0 and 2, which match the first sense of the word (e.g., new virus strain), shrink. This behavior underlines the evolution of the concerns related to the pandemic in the newspapers.



#	Keywords
0	strain coronavirus, new strain, city wuhan, novel strain, strain virus, chinese city
1	strain health, strain resources, stream, network infrastructure, international resources, likely strain
2	new strain, acute respiratory, 2019 ncov, respiratory syndrome, severe acute, identified humans
3	financial strain, feeling strain, strain coronavirus, economic strain, signs strain, strain said
4	ease strain, putting strain, strain health, reduce strain, care system, strain hospitals

Figure 2: Cluster distributions per month and top keywords for each cluster for word *strain*.

7 Conclusion

We proposed a scalable and interpretable method for word usage change detection, which outperforms the non-scalable contextual embeddings-based methods by a large margin. The new method also allows completely data-driven analysis of word sense dynamic in large corpora, which was impossible to conduct with unscalable methods. This opens new opportunities in both language change studies and text stream monitoring tasks. In this paper we focused on the latter application by analysing a large corpus of COVID-19 related news.

The method is outperformed by the state-of-the-art SGNS+OP+CD method. We hypothesise that this can be connected with the fact that the sentences in all but one evaluation corpus (COHA) are shuffled, meaning that BERT models cannot leverage the usual sequence of 512 tokens as a context, but are limited to the number of tokens in the sentence. We will explore this hypothesis in the future.

Despite achieving lower performance than the SGNS+OP+CD method, we nevertheless argue that our method offers a more fine-grained interpretation than methods based on non-contextual embeddings, since it accounts for the fact that words can have multiple meanings. The cluster-based technique returns a degree of change and a set of sentence clusters for each word in the corpus, roughly corresponding to word senses or particular usages. For this reason, the approach can be used for detection of new word usages and for tracing how these usages disappear, as we have shown in Section 6. Even more, word usages and their distributions over time could be linked with real-word events by labeling sentence clusters with a set of cluster-specific keywords.

Overall, we observe a large disparity between results on different evaluation corpora. This is in line with the results of the Semeval 2020 task 1 (Schlechtweg et al., 2020), where none of the best-performing methods was able to achieve the best result on all corpora. In practice, different methods focus on different aspects of word usage change: Averaging and SGNS+OP+CD focus on average variation of word usage, hiding the intra-period diversity. When it comes to clustering, JSD-based method detects the appearance or disappearance of a given usage, even a minor one. The WD-based method, using information from both the cluster distribution and the embeddings vectors, represents a compromise between the averaging and the JSD-based methods.

In this paper we follow the general approach in semantic shift detection literature and apply our analysis on the raw text. However, our results demonstrate that at least news monitoring applications would benefit from the application of the traditional text processing pipeline, in particular the extraction of named entities and dates. This will be addressed in the future work.

Acknowledgements

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA), the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581), and Project Development of Slovene in the Digital Environment (RSDO).

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Felix Hennig and Steven Wilson. 2020. Diachronic embeddings for people in the news. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 173–183.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics.

J. Lin. 2006. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020a. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020b. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4811—4819.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. *CoRR*, abs/2001.03216.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Justin Solomon. 2018. Optimal transport on discrete domains.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. In *Eleventh international AAAI conference on web and social media*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, 1811.06278.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in neural information processing systems*, pages 9412–9423.

Appendix B: EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+

EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+

Matej Martinc
Jožef Stefan Institute
Jamova 39, Ljubljana
matej.martinc@ijs.si

Nina Perger
Faculty of Social Sciences
Kardeljeva ploščad 5, Ljubljana
nina.perger@fdv.uni-lj.si

Andraž Pelicon
Jožef Stefan Institute
Jamova 39, Ljubljana
andraz.pelicon@@ijs.si

Matej Ulčar
Faculty of Computer Science
Večna pot 113, Ljubljana
matej.ulcar@fri.uni-lj.si

Andreja Vezovnik
Faculty of Social Sciences
Kardeljeva ploščad 5, Ljubljana
andreja.vezovnik@fdv.uni-lj.si

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

Abstract

We conduct automatic sentiment and viewpoint analysis of the newly created Slovenian news corpus containing articles related to the topic of LGBTIQ+ by employing the state-of-the-art news sentiment classifier and a system for semantic change detection. The focus is on the differences in reporting between quality news media with long tradition and news media with financial and political connections to SDS, a Slovene right-wing political party. The results suggest that political affiliation of the media can affect the sentiment distribution of articles and the framing of specific LGBTIQ+ specific topics, such as same-sex marriage.

1 Introduction

Quantitative content analysis of news related to LGBTIQ+ in general, and specifically, to marriage equality debates show that distinctions can be drawn between those media articles that express positive, neutral or negative stance towards same-sex marriage. Those media articles that express positive stance are grounded in human rights/civil equality discourses and access to benefits (Zheng and Chan, 2020; Colistra and Johnson, 2019; Paterson and Coffey-Glover, 2018), and frame marriage equality as an inevitable path towards equality, as a civil right issue that would reduce existing prejudices and discrimination, and protect threatened LGBTIQ+ minority (Zheng and Chan, 2020).

For media articles that express negative stance towards marriage equality, distinctive discursive elements are present, such as "equal, but separate" (marriage equality should be implemented, but differentiating labels should be kept in the name of protecting the institute of marriage) (Kania, 2020; Zheng and Chan, 2020; Paterson and Coffey-Glover, 2018), and reference procreation/welfare of children (Kania, 2020; Zheng and Chan, 2020), public objection (Kania, 2020) and church – state opposition (Paterson and Coffey-Glover, 2018).

The related work also shows that the differences between "liberal" and "conservative" arguments are not emphasised, mostly because both sides refer to each other's arguments, if only to negate them; yet, political orientation can be identified through the tone of the article (Zheng and Chan, 2020).

When it comes to methods employed for automatic analysis of the LGBTIQ+ topic, most recent approaches rely on embeddings. Hamilton et al. (2016) employed embeddings to research how words (among them also word *gay*) change meaning through time. They built static embedding models for each time slice of the corpus and then make these representations comparable by employing *vector space alignment* by optimising a geometric transformation. This research was recently expanded by (Shi and Lei, 2020), who employed embeddings to explore semantic shifts of six descriptive LGBTIQ+ words from the 1860s to the 2000s: *homosexual, lesbian, gay, bisexual, transgender*, and *queer*.

There are also several general news analysis techniques that can be employed for the task at hand. Azarbonyad et al. (2017) developed a system for semantic shift detection for viewpoint analysis of political and media discourse. A recent study by Spinde et al. (2021) tried to identify biased terms in news articles by comparing news media outlet specific word embeddings. On the other hand, Pelicon et al. (2020) developed a system for analysing the sentiment of news media articles.

While the above described analyses in a large majority of cases covered news in English speaking countries, in this research, we expand the quantitative analysis to Slovenian news, in order to determine whether attitudes towards LGBTIQ+ differs in different cultural environments. We created a corpus of LGBTIQ+ related news and conducted an

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 121–126

automatic analysis of its content covering several aspects:

- Sentiment of news reporting, where we focused on the differences in reporting between well established media with long tradition of news reporting and more recently established media characterised by their financial and political connections to the Slovene conservative political party SDS.
- Usage of words, where we tried to identify the words that are used differently in different news sources and would indicate the difference in the prevailing discourse on the topic of LGBTIQ+ in the specific liberal and conservative media.

The research was performed in the scope of the EMBEDDIA Hackashop (Hackaton track) at EACL 2021 and employs several of the proposed resources and tools (Pollak et al., 2021).

2 Methodology

For **sentiment analysis** we used a multilingual news sentiment analysis tool. The tool was trained using a two-step approach, described in Pelicon et al. (2020). For training, a corpus of sentiment-labeled news articles in Slovenian was used (Bucar et al., 2018) with news covering predominantly the financial and political domains. This model was subsequently applied to the LGBTIQ+ corpus where each news article was labeled with one of the sentiment labels, namely negative, neutral or positive. This allowed us to generate a sentiment distribution of articles for each media source in the corpus.

For **word usage viewpoints analysis**, we applied a system originally employed for diachronic shift detection (Martinc et al., 2020b). Our word usage detection pipeline follows the procedure proposed in the previous work (Martinc et al., 2020a,b; Giulianelli et al., 2020): the created LGBTIQ+ corpus is split into two slices containing news from different news source according to procedure described in Section 3. Next, the corpus is lemmatized, using the Stanza library (Qi et al., 2020), and lowercased. For each lemma that appears more than 100 times in each slice and is not considered a stopword, we generate a slice specific set of contextual embeddings using BERT (Devlin et al., 2019) pretrained on the Slovenian, Croatian and English texts (Ulčar and Robnik-Šikonja, 2020). These representations are clustered using k-means and the derived cluster distributions are compared across slices by employing Wasserstein distance (Solomon, 2018). It is assumed that the ranking resembles a relative degree of usage change, therefore words are ranked according to the distance.

Once the most changed words are identified, the next step is to understand how their usage differs in the distinct corpus slices. The hypothesis is that specific clusters of BERT embeddings resemble specific word usages of a specific word. The problem is that these clusters may consist of several hundreds or even thousands of word usages, i.e. sentences, therefore manual inspection of these usages would be time-consuming. For this reason, we extract the most discriminating unigrams, bigrams, trigrams and fourgrams for each cluster using the following procedure: we compute the term frequency - inverse document frequency (tf-idf) score of each n-gram and the n-grams appearing in more than 80% of the clusters are excluded to ensure that the selected keywords are the most discriminant. This gives us a ranked list of keywords for each cluster and the top-ranked keywords (according to tf-idf) are used for the interpretation of the cluster.

3 Experiments

3.1 Dataset

The corpus was collected from the Event registry (Leban et al., 2014) dataset by searching for Slovenian articles from 2014 to (including) 2020, containing any of the manually defined 125 keywords (83 unigrams and 42 bigrams) and their inflected forms connected to the subject of LGBTIQ+. The resulting corpus contains news articles on the LGBTIQ+ topic from 23 media sources. The corpus statistics are described in Table 1. Out of this corpus, we extracted a subcorpus appropriate for the viewpoint analysis. The subcorpus we used included the following online news media: Delo, Večer, Dnevnik, Nova24TV, Tednik Demokracija and PortalPolitikis. The sources were divided into two groups. The first group, namely Delo, Večer and Dnevnik represent the category of daily quality news media that are published online and in print with a long tradition in the Slovene media landscape. These three media are relatively highly trusted by readers and have the highest readership amongst Slovene dailies. The second group of news media - namely, Nova24TV, Ted-

Source	Num. articles	Num. words
MMC RTV Slovenija	1790	1,555,977
Delo	1194	1,064,615
Nova24TV	844	683,336
Večer	667	552,195
24ur.com	661	313,794
Dnevnik	592	262,482
Siol.net Novice	549	460,561
Slovenske novice	501	236,516
Svet24	430	286,429
Mladina	394	275,506
Tednik Demokracija	361	350,742
Domovina	327	283,478
Primorske novice	255	183,624
Druzina.si	253	149,761
Vestnik	242	263,737
Časnik.si - Spletni magazin z mero	239	280,339
Žurnal24	172	79,953
PortalPolitikis	157	111,683
Revija Reporter	102	62,429
Gorenjski Glas	97	92,751
Onaplus	79	104,343
Športni Dnevnik Ekipa	67	33,936
Cosmopolitan Slovenija	57	71,538

Table 1: LGBTIQ+ corpus statistics.

nik Demokracija and PortalPolitikis have been established more recently and are characterised by their financial and political connections to the Slovene right-wing/conservative political party SDS (Slovenska demokratska stranka) and the Roman Catholic Church.

3.2 Sentiment Analysis

Figure 1 presents sentiment distribution across articles for each specific news media, arranged from left to right according to the share of articles with negative sentiment. Note that all three media houses selected for the viewpoint analysis (Nova24TV, Tednik, Demokracija and PortalPolitikis) because of their financial and political connections to the Slovene right-wing/conservative political party SDS produce more news articles with negative sentiment on the topic of LGBTIQ+ than the mainstream media with the long tradition (Delo, Dnevnik, Večer). The source with the most negative content about LGBTIQ+ is Revija Reporter, which is in most media analyses positioned in the right-wing ideological spectrum[1] (Milosavljević, 2016; Milosavljević and Biljak Gerjevič, 2020). On the other side the source with the smallest share of negative news is Primorske novice, a politically independent daily regional quality news media published online and in print with a long tradition in the regional media landscape. Nevertheless, not all conservative media are characterized by a more negative reporting about the LGBTIQ+ topic. For example, the source with the second lowest share of negative news is Druzina.si, which is strongly connected to Roman Catholic Church.

[1] https://podcrto.si/mediji-martina-odlazka-1-del-nepregledna-mreza-radiev-tiskovin-televizije/

3.3 Viewpoint Analysis

The viewpoint analysis was conducted by finding words, whose usage varies the most in the two groups of media sources selected for the analysis (i.e. Delo, Dnevnik, Večer vs. Nova24TV, Tednik Demokracija and PortalPolitiks). The 10 most changed words are presented in Table 2. The word that changed the most was globok (deep), for which our system for interpretation of the change revealed that it was selected due to frequent mentions of *deep state* in the media with connections to political right. The context of *deep state* is interesting, since it is a very frequently used interpretative frame by this group of media sources, regardless of the specific topic. Here it indicates the framing of the LGBTIQ+ questions as part of a political agenda driven by the left-wing politics. The second word roman (novel) was selected because it appears in two contexts: as a novel and also as a constituent word in a name of the Slovenian LGBTIQ+ activist, Roman Kuhar. While the third word, *video*, is a corpus artefact that offers little insight into the attitude towards LGBTIQ+, the fourth word, *razmerje* (relationship), has a direct connection to some of the most dividing LGBTIQ+ topics, such as gay marriage, therefore for this word we provide a more detailed analysis. Figure 2 presents cluster distributions per two media groups and top 5 (translated) keywords for each cluster for word *razmerje(relationship)*. The main difference between the two distributions can be observed when it comes to mention of relationship in the context of family and marriage (see the red cluster), which present a large cluster of usages in the mainstream media but a rather small cluster in the right-wing

1	globok(deep)	6	napaka(mistake)
2	roman(novel)	7	nadaljevanje(continuation)
3	video	8	lanski(last year)
4	razmerje(relationship)	9	kriza(crisis)
5	teorija(theory)	10	pogledat(look)

Table 2: Top 10 most changed words (and their English translations) in the corpus according to Wasserstein distance between k-means ($k = 5$) cluster distributions in distinct chunks of the corpus.

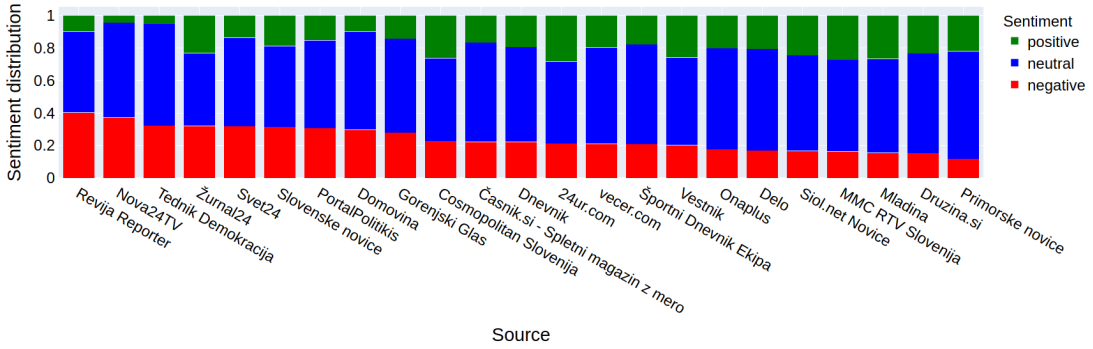
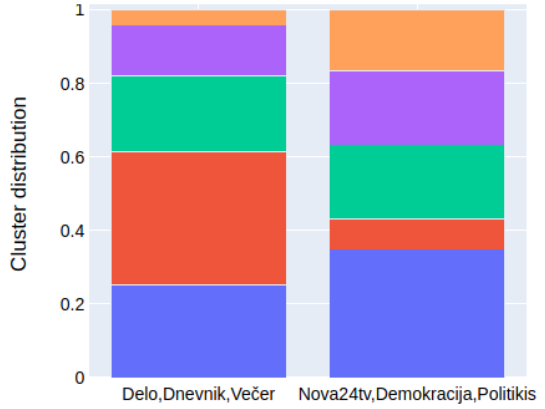


Figure 1: Sentiment distribution for each source in the LGBTIQ+ corpus.



#	Keywords
	partnership, homosexual relationship, polygamous relationship, intimate relationship, problem
	family relationship, end work relationship, woman, field, less
	power in the relationship, relationship between people, percentage, man and woman, story
	family relationship, marriage, children, Ljubljana, coalition for children
	woman, love relationship, work relationship, film, meet

Figure 2: Cluster distributions per two media groups and top 5 translated keywords for each cluster for word *razmerje(relationship)*.

media. On the other hand, relationship is in these media mentioned a lot more in the context of partnership, homosexuality and polygamy (see the orange cluster). The other three clusters (i.e., usages) have a rather strong presence in both media groups.

4 Conclusions

We conducted a content analysis of the Slovenian news corpus containing articles related to the topic of LGBTIQ+. The sentiment analysis study shows that there are some differences in the sentiment of reporting about LGBTIQ+ between two distinct groups of media and that the three media houses connected to political right tend to cover the subject in a more negative manner. This supports the thesis by Zheng and Chan (2020), who suggested that political orientation can be identified through the tone of the article. Nevertheless, the obtained results should be interpreted with the grain of caution, since the sentiment classifier we employed cannot distinguish whether it is the stance expressed towards the LGBTIQ+ community, or is it rather the event on which the article is reporting, that is positive or negative (e.g., an attack on the LGBTIQ+ activist). The distinction between these two "types" of sentiment will be analysed in the future work.

The viewpoint analysis suggests that the usage of some specific words has been adapted in order to express specific ideological point of view of the media. For example, the analysis of the word *relationship* suggests that the more conservative media more likely frame LGBTIQ+ relationships as a *partnership* of two homosexual (or even polygamous) partners. On the other hand, they rarely consider LGBTIQ+ relationships as family or talk about marriage.

In the future we plan to conduct topic analysis of the corpus in order to identify the most common LGBTIQ+ related topics covered by the news media. We will also employ embeddings to research relations between LGBTIQ+ specific words.

Acknowledgments

This work was supported by the Slovenian Research Agency (ARRS) grants for the core programme Knowledge technologies (P2-0103), the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581), as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Joze Bucar, M. Znidarsic, and J. Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52:895–919.

Rita Colistra and Chelsea Betts Johnson. 2019. Framing the legalization of marriage for same-sex couples: An examination of news coverage surrounding the us supreme court's landmark decision. *Journal of homosexuality*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Ursula Kania. 2020. Marriage for all ('ehe fuer alle')?! a corpus-assisted discourse analysis of the marriage equality debate in germany. *Critical Discourse Studies*, 17(2):138–155.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 107–110.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020a. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Discovery team at semeval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73.

Marko Milosavljević. 2016. Media pluralism monitor 2016 monitoring risks for media pluralism in the EU and beyond - country report: Slovenia.

Marko Milosavljević and Romana Biljak Gerjevič. 2020. Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the European Union, Albania and Turkey in the years 2018-2019 - country report: Slovenia.

Laura L Paterson and Laura Coffey-Glover. 2018. Discourses of marriage in same-sex marriage debates in the uk press 2011–2014. *Journal of Language and Sexuality*, 7(2):175–204.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Yaqian Shi and Lei Lei. 2020. The evolution of lgbt labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, 36(4):33–39.

Justin Solomon. 2018. Optimal transport on discrete domains.

Timo Spinde, Lada Rudnitckaia, and Felix Hamborg. 2021. Identification of biased terms in news articles

by comparison of outlet-specific word embeddings. In *Proceedings of the 16th International Conference (iConference 2021). Springer Nature, Virtual Event China*.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.

Yue Zheng and Lik Sam Chan. 2020. Framing same-sex marriage in us liberal and conservative newspapers from 2004 to 2016: Changes in issue attributes, organizing themes, and story tones. *The Social Science Journal*, pages 1–13.

Appendix C: Zero-Shot Learning for Cross-Lingual News Sentiment Classification

applied sciences

MDPI

Article

Zero-Shot Learning for Cross-Lingual News Sentiment Classification

Andraž Pelicon [1,2,*], Marko Pranjić [2,3], Dragana Miljković [1], Blaž Škrlj [1,2] and Senja Pollak [1,*]

1 Jožef Stefan Institute, 1000 Ljubljana, Slovenia; dragana.miljkovic@ijs.si (D.M.); blaz.skrlj@ijs.si (B.Š.)
2 Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia; marko.pranjic@styria.ai
3 Trikoder d.o.o., 10010 Zagreb, Croatia
* Correspondence: Andraz.Pelicon@ijs.si (A.P.); senja.pollak@ijs.si (S.P.)

Received: 31 July 2020; Accepted: 25 August 2020; Published: 29 August 2020

check for updates

Abstract: In this paper, we address the task of zero-shot cross-lingual news sentiment classification. Given the annotated dataset of positive, neutral, and negative news in Slovene, the aim is to develop a news classification system that assigns the sentiment category not only to Slovene news, but to news in another language without any training data required. Our system is based on the multilingual BERTmodel, while we test different approaches for handling long documents and propose a novel technique for sentiment enrichment of the BERT model as an intermediate training step. With the proposed approach, we achieve state-of-the-art performance on the sentiment analysis task on Slovenian news. We evaluate the zero-shot cross-lingual capabilities of our system on a novel news sentiment test set in Croatian. The results show that the cross-lingual approach also largely outperforms the majority classifier, as well as all settings without sentiment enrichment in pre-training.

Keywords: sentiment analysis; zero-shot learning; news analysis; cross-lingual classification; multilingual transformers

1. Introduction

Sentiment analysis is one of the most popular applications of natural language processing (NLP) and has found many areas of applications in customers' product reviews, survey textual responses, social media, etc. It analyzes users' opinions on various topics, such as politics, health, education, etc. In sentiment analysis, the goal is to analyze the author's sentiments, attitudes, emotions, and opinions [1]. Traditionally, such analysis was performed towards a specific entity that appears in the text [2]. A less researched, but nevertheless prominent field of research in sentiment analysis is to shift the focus from analyzing sentiment towards a specific target to analyzing the intrinsic mood of the text itself. Several works try to model feelings (positive, negative, or neutral) that readers feel while reading a certain piece of text, especially news [3,4]. In Van de Kauter et al. [5], the authors claimed that the news production directly affects the stock market as the prevalence of positive news boosts its growth and the prevalence of negative news impedes it. In the context of news media analytics, the sentiment of news articles has been used also as an important feature in identifying fake news [6] and biases in the media [7]. Rambaccussing and Kwiatkowski [8] explored the change in sentiment of news articles from major U.K. newspapers with respect to current economic conditions. Bowden et al. [9] took a step further and tried to improve the forecasting of three economic variables, inflation, output growth, and unemployment, via sentiment modeling. They concluded that, using sentiment analysis, out of the three variables observed, the forecasting can be effectively improved for unemployment.

In the last year, the use of pre-trained Transformer models has become standard practice in modeling text classification tasks. Among the first such models was the BERT (Bidirectional Encoder Representations from Transformers) model developed by [10], which achieved state-of-the-art performance on several benchmark NLP tasks, as well as in real-world applications, e.g., Google search engine [11] and chatbots [12]. The initial model was however pre-trained only on English corpora and could consequently be used only for modeling textual data in the English language. A new version of the BERT model, titled multilingual BERT or mBERT, soon followed. This model was pre-trained on unlabeled data in 104 languages with the largest Wikipedias using a joint vocabulary. Several studies noted the ability of the mBERT model to work well in multilingual and cross-lingual contexts even though it was trained without an explicit cross-lingual objective and with no aligned data [13,14].

In the context of sentiment analysis of news articles, we however identified two potential drawbacks of the mBERT model. The first is that the model accepts the inputs of a fixed length where the length is determined by the length of the context window, i.e., the maximum length of the input sequence during the pre-training phase. Since the training becomes computationally more expensive with the size of the context window, several standard implementations of the mBERT model have the context window set to a maximum length [15]. The standard solution for longer documents is therefore to cut the inputs to the length of the context window [16]. This method however potentially causes the loss of important information that could be present in the later parts of the document. Another potential drawback is that the input representations produced by the Transformer models may encode only a small amount of sentiment information. The pre-training objectives, namely the masked language modeling and next sentence prediction, are designed to focus on encoding general syntactic and certain semantic features of a language. The only explicit sentiment signal the models get is during the fine-tuning phase, when the models are generally trained on a much smaller amount of data.

The paper presents the advances achieved in the scope of the European project H2020 EMBEDDIA (www.embeddia.eu, duration 2019–2021), which focuses on the development of cross-lingual techniques to transfer natural language processing tools to less-resourced European languages with applications to the news media industry. In this paper, we present our approach to cross-lingual news sentiment analysis, where given an available sentiment-annotated dataset of news in Slovene [3], we propose a news sentiment classification model for other languages. In this paper, we focus on Croatian, where the news dataset is provided by 24sata, one of the leading portals in Croatia, and was labeled with the same sentiment annotation scheme as the Slovenian dataset in order to allow comparison in a zero-shot learning setting where no annotations in the target language are expected.

We identify three main contributions of this paper focusing mainly on the cross-lingual zero-shot learning setting. First, we gathered a sentiment-annotated corpus of Croatian news, where the annotation guidelines follow the annotation scheme of the Slovenian sentiment-annotated news dataset [3], therefore enabling cross-lingual zero-shot learning sentiment evaluation. Second, we tested several document representation techniques to overcome one of the shortcomings of the BERT models of not being capable of efficiently processing longer text documents. Last, but not least, we propose a novel intermediate training step to directly enrich the BERT model with sentiment information in order to produce input representations of better quality for sentiment classification tasks. These representations were then tested both in a monolingual setting, as well as in the zero-shot cross-lingual setting, where the model was tested on a different language without any additional target language training. Our experiments show that these representations improve the results in the monolingual setting and achieve a substantially better result than the majority baseline classifier in the cross-lingual setting.

The article is structured as follows. In Section 2, we first present the related work upon which our study builds. In Section 3, we present two datasets of news articles that are manually labeled in terms of sentiment: the existing Slovenian dataset [3] and the newly constructed Croatian test set. Section 4, where we present the methodology, is followed by Section 5, explaining the experimental

setup, with the training regime applied and the evaluation method. Section 6 presents the results of the experiments and discusses their impact, which is followed by qualitative inspection of the models in Section 7. Section 9 presents the conclusions of this work and ideas for future research.

2. Related Work

Traditionally, sentiment analysis was modeled through the use of classical machine learning methods, where especially learners such as support vector machines combined with the TF-IDF text representations proved to be widely successful [17,18]. Lately, however, deep neural networks have become more frequent for sentiment analysis and started outperforming the classical approaches. Mansar et al. [19] used convolutional neural networks (CNN), a variant of neural networks, which are heavily utilized for computer vision. With the help of the convolutional layer, they acquired word-level representations of individual news articles from the learning corpus and combined them with the sentiment score of the individual article, which was obtained with a simple, rule-based model. The attributes were used as input to the fully connected NN. Their model showed the best performance on the SemEval2017 challenge (Task 5, Subtask 2). Moore and Rayson [20] used two models for analyzing sentiment in financial news titles, a support vector machine and a bidirectional LSTM (Long-Short Term Memory) neural network. They reported the LSTM neural network to outperform the SVM modelsby 4–6%.

Several recent works also explored the problem of cross-lingual sentiment analysis. One of the earlier studies [21] employed machine translation to translate a large corpus of sentiment-annotated English training data for the development of a Chinese sentiment classifier. These translated data were then used in addition to the original Chinese data to train an SVM-based classifier. While machine translation can be a good solution for cross-lingual modeling, a quality machine translation system for a particular language pair may not exist or may be expensive to train. Furthermore, machine learning systems struggle with distant language pairs [22]. Zhou et al. [23] developed a cross-lingual English-Chinese attention-based neural architecture for sentiment classification. It utilizes a two-level hierarchical attention mechanism. The first layer of the model encodes each sentence separately by finding the most informative words. Then, the second layer produces the final document representation from lower-level sentence representations. The downside of their work is that the model uses aligned data in two languages, which are not readily available for every language pair. Ref. [24] proposed a representation learning method that utilizes emojis as an instrument to learn language-independent sentiment-aware text representations. The approach is however limited to text types where emojis regularly appear. The cross-lingual sentiment classification approaches presented above also do not address news analysis, but focus on shorter social media texts, where there is no need for adaptation to longer text sequences and they do not leverage cross-lingual Transformer models, such as mBERT, that have been recently introduced as the state-of-the-art for cross-lingual classification tasks. In this paper, we will bridge this gap by proposing a novel approach where we not only leverage standard transfer learning where pretrained language models are fine-tuned for specific classification tasks (in the same or another language), but introduce a novel intermediate training step for sentiment enrichment of BERT models.

The need for labeled data is seen as one of the main obstacles in developing robust cross-lingual systems for natural language processing, especially for low-resource languages. For this reason, research has been focused lately on models that can work in a zero-shot setting, i.e., without being explicitly trained on data from the target language or domain. This training paradigm has been utilized with great effect for several popular NLP problems, such as cross-lingual document retrieval [25], sequence labeling [26], cross-lingual dependency parsing [27], and reading comprehension [28]. More specific to classification tasks, Ye et al. [29] developed a reinforcement learning framework for cross-task text classification, which was tested also on the problem of sentiment classification in a monolingual setting. Jebbara and Cimiano [30] developed models for cross-lingual opinion target extraction, which were tested in a zero-shot setting, similar to ours. Their approaches rely on the

alignment of static monolingual embeddings into the shared vector space for input representation. Fei and Li [31] trained a multi-view cross-lingual sentiment classifier based on the encoder-decoder architecture used for unsupervised machine translation. Their systems showed state-of-the-art performance on several benchmark datasets. The difference from our work is that the datasets used are all product review datasets, which contain considerably shorter texts. Furthermore, as described in Section 1, product reviews contain the target of the modeled sentiment in the text, while news articles generally do not, which makes the two problems different on a more fundamental level.

Novel research has also been done on better input text representation techniques for classification tasks. Tan et al. [32] proposed a clustering method for words based on their latent semantics. The vectors composing the same clusters were then aggregated together into cluster vectors. The final set of cluster vectors was then used as the final text representations. This novel text representation technique showed improvement on five different datasets. Pappagari et al. [33] proposed a modification to the BERT model for long document classification in a monolingual setting. They utilized a segmentation approach to divide the input text sequences into several subsequences. For each subsequence, they obtained a feature vector from the Transformer, which they then aggregated into one vector by applying another LSTM- or Transformer-based model over it. This work has inspired part of our current research for obtaining better Transformer-based representation of long text sequences. Ref. [34] recently presented a Transformer architecture, which is able to produce input representations from long documents in an efficient manner. However, the model they produced based on this architecture was pre-trained only on English data.

3. Datasets

In this section, we present in detail the two datasets of sentiment-labeled news that were used in this experiment.

3.1. SentiNews Dataset in Slovene

We used the publicly available SentiNews dataset (available at https://www.clarin.si/repository/xmlui/handle/11356/1110) [3], which is a manually sentiment-annotated Slovenian news corpus. The dataset contains 10,427 news texts mainly from the economic, financial, and political domains from Slovenian news portals (www.24ur.com, www.dnevnik.si, www.finance.si, www.rtvslo.si, www.zurnal24.si), which were published between 1 September 2007 and 31 December 2013. The texts were annotated by two to six annotators using the five-level Likert scale on three levels of granularity, i.e., on the document, paragraph, and sentence level. The dataset contains information about average sentiment, standard deviation, and sentiment category, which correspond to the sentiment allocation according to the average sentiment score. The dataset statistics are:

- 10,427 documents;
- 89,999 paragraphs;
- 168,899 sentences.

For our news classification experiments, we used the document-level annotations, with 10,427 news articles and an imbalanced distribution of 3337 (32%) negative, 5425 (52%) neutral, and 1665 (16%) positive news, where the sentiment category corresponds to the sentiment allocation according to the average sentiment score. For intermediate training, we also leveraged paragraph-level annotations.

3.2. Croatian Sentiment Dataset

The Croatian dataset was annotated in the scope of project EMBEDDIA and for the purposes of testing cross-lingual classification; therefore, the annotation procedure fully matched the Slovenian dataset [3].

The data came from 24sata, one of the leading media companies in Croatia with the highest circulation newspaper. The 24sata news portal is one of the most visited websites in Croatia, and it

consists of a portal with daily news and several smaller portals covering news from specific topics such as automotive news, health, culinary content, and lifestyle advice. Portals included in the dataset are www.24sata.hr (daily news content, the majority of the dataset), as well as miss7.24sata.hr, autostart.24sata.hr, joomboos.24sata.hr, miss7mama.24sata.hr, miss7zdrava.24sata.hr, www.express.hr, and gastro.24sata.hr.

The dataset statistics are:

- 2025 documents;
- 12,032 paragraphs;
- 25,074 sentences.

As in [3], the annotators chose the sentiment score on the Likert [35] scale (corresponding to the question: Did this news evoke very positive/positive/neutral/negative/very negative feelings?), but for the final dataset, the average annotations were then three classes (positive, negative, and neutral). Annotations were done on three levels: document, paragraph, and sentence level. The distribution of positive, negative and neutral news texts of the document-level annotations used in this study is as follows: 303 (15.1%) positive, 439 (21.5%) negative, and 1283 (63.4%) neutral. They will be made available under a CC license upon acceptance of the paper. More details about inter-annotator agreement and annotation procedure are available in the Appendix A of this paper.

As one of the contributions of this paper is the evaluation of representation learning for long articles, we also provide the statistics of both datasets in terms of length. Table 1 compares the Slovenian and Croatian news datasets in terms of the length of annotated articles. It presents the average number of tokens per article, as well as the length of the longest and shortest articles in the respective datasets. We present the lengths in terms of the standard tokenization procedure where each word and punctuation mark counts as a separate token. However, the BERT model uses a different form of tokenization, namely the WordPiece tokenization [36]. Using this tokenization process, each word is broken into word pieces, which form the vocabulary of the tokenizer. The vocabulary is obtained using a data-driven approach: given a training corpus G and a number of word pieces D, the task is to select D word pieces such that the segmented corpus G contains as much unsegmented words as possible. The selected word pieces then form the vocabulary of the tokenizer. This approach is proven to handle the out-of-vocabulary words better than standard tokenization procedures. Since the inputs to the BERT model have to be tokenized according to this algorithm in order for the model to properly learn, we present the length statistics in terms of BERT's WordPiece tokenization model as well in the column "BERT tokens". We may observe that the average length of the articles in both datasets is relatively long in terms of the BERT tokens. Especially in the Slovenian dataset, which is used for training in this experiment, the average length of an article surpasses the maximum window size of the BERT model, which is set to 512 tokens in the implementation we are using for this work.

Table 1. Length of the articles in the Slovenian and Croatian datasets in terms of the number of tokens. The row "Tokens" presents the length in terms of the standard tokenization procedure, and the row "BERTtokens" presents the length of the articles in terms of BERT's WordPiece tokenization.

	Slovenian			Croatian		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**
Tokens	10	2833	350	155	515	273
BERT tokens	19	4961	648	256	816	456

4. Methodology

We tested two approaches, one focusing on techniques for long document representation and the second one on improving the performance on the sentiment analysis task through intermediate pre-training.

In this work, we model sentiment in news articles, which are frequently longer than the BERT context windows, as discussed in Section 1. Therefore, in our first approach, we experiment with several methods for representing longer documents.

The second approach, presented in Section 4.4, proposes a novel technique for sentiment enrichment of mBERT. In standard BERT architectures, the pre-training phase of BERT consists of masked language modeling and next sentence prediction tasks, which are robust, but not necessarily relevant for sentiment classification, as discussed in Section 1. Therefore, we add an intermediate training step where, aside from masked language modeling, the sentiment classification is used as a learning objective. This model is then used for final fine-tuning. The role of intermediate training for BERT is still unexplored in NLP, with some initial experiments presented in [37].

4.1. Beginning of the Document

In the first experimental setting, we produced the document representations by using only the beginning part of the document. We first tokenized the document with the pre-trained multilingual BERT tokenizer. We then took the sequence of 512 tokens from the beginning of the document and fed them to the BERT language model. As proposed in Devlin et al. [10], we used the representation of the [CLS] token produced by the language model as the document representation. The [CLS] token is a special token prepended to every input of the BERT model, which, after fine-tuning, is used to represent the input sequence for classification tasks. We then sent this representation to the classification head composed of a single linear layer. This experiment mimics the usual usage of the BERT pre-trained models for text classification tasks and is included in this work for better benchmarking of other proposed text representation methods.

4.2. Beginning and End of the Document

For the second setting, we tried to produce the document representations by using the beginning and end of the document. The length of the input sequence was retained at 512 tokens. For sequences longer than 512 tokens after tokenization, we took 256 tokens from the beginning of the text and 256 tokens from the end of the text and concatenated them. We then fed the sequence to the BERT language model and used the [CLS] token vector from the last layer as the document representation. This document representation was then fed to the classification head composed of a linear layer.

4.3. Using Sequences from Every Part of the Document

In the third setting, we tried to compose our document representation by using information in the whole document.

For the language model fine-tuning phase, we tokenized each document and broke it into sequences of 512 tokens. We then used a sliding window that moved over all the subsequences in the order they appeared in the original sequence. Each subsequent window would overlap the first fifty tokens from the previous window. This way, we hoped our model would capture the relationships across sentence boundaries. We attached the document sentiment label to each of the subsequences from the same document. Such an oversampled dataset was then used to fine-tune the multilingual BERT language model with the attached linear layer for classification. This method is graphically presented in Figure 1.

After finetuning we again prepared each document in the dataset as described above and sent every subsequence of a particular document to the fine-tuned BERT model. We extracted the [CLS] vector representations from the last layer and combined them into a final document representation. This approach is inspired by the work of Pappagari et al. [33]. The main difference of our study is in the way the subsequence representations are merged into a document representation. In this work, we tested three different ways of combining the output vector representations into the final document representation.

- Using the most informative subsequence representation:

  In this approach, we tried to identify the most informative subsequence for the task at hand. As the BERT language model was fine-tuned on the sentiment classification task, we assumed some notion of the importance of different parts of the text was encoded directly into the vector representations. Using this line of thought, we defined the most informative subsequence as the subsequence with the highest euclidean vector norm. Formally, from the set of ordered subsequence representations: $S = \{x_1, x_2, \ldots x_n\}$ we chose: $x = argmax(||x||_2 : x \in S)$. We then used only this representation as the final vector representation and discarded the rest. The document representation is then sent into a two-layer fully connected neural network, which produces the final predictions.

- Averaging the representations of all subsequences:

  As the first approach is based on a strong assumption and it does not actually utilize the data from the whole document, here we combine all the vector representations of subsequences into one final document representation. We used a relatively naive approach of simply averaging all the vector representations to produce the final document embedding. The document representation is then sent into a two-layer fully connected neural network, which produces the final predictions.

- Using convolutional layers:

  In this approach, we extracted the most informative parts of the document with the use of 1D convolutional neural layers. We used a convolutional filter of size 2 with stride 2 that runs over the produced subsequence representations. This way, the convolutional filter processes the subsequences in pairs and extracts the most informative features from each pair of subsequences from each part of the document. Since we have documents of variable lengths that may be represented by a variable number of subsequences, all the representations were padded with zero vectors up to the maximum length of 6. We used 128 filters to produce 128 feature maps. We then mapped these maps to a final 128-dimensional document vector representation using a max pooling operation. The final embedding is then sent into a linear layer that produces the classification.

The advantage of the first two mapping operations is that, in comparison to the methods proposed in Pappagari et al. [33], they are more computationally efficient as we need to perform simple vector norm and averaging calculations to produce the final document representations. The third mapping operations uses a convolutional layer to map the different subsequences into one document representation. The convolutional networks have proven in the past to be competitive with other text-processing approaches in NLP [38]; therefore, our approach presents an alternative to the LSTM and Transformer-based sequence aggregation.

4.4. Sentiment Enrichment of the mBERT Model

In this approach, the aim is to to induce sentiment information directly into the vectorized document representations that are produced by the multilingual BERT model. To do so, we added an intermediate training step for the mBERT model before the fine-tuning phase. The intermediate training phase consists of jointly training the model on two tasks. The first task we used was the masked language modeling task as described in the original paper by Devlin et al. [10]. We left this task unchanged in hopes that the model would better capture the syntactic patterns of our training language and domain.

For the second task, we used the sentiment classification task, which mirrors the fine-tuning task, but is trained using a different set of labeled data. With this task, we tried to additionally constrain the model to learn sentiment-related information before the actual fine-tuning phase. The task was

formally modeled as a standard classification task where we tried to learn a predictor that would map the documents to a discrete number of classes:

$$\gamma : x \rightarrow C$$

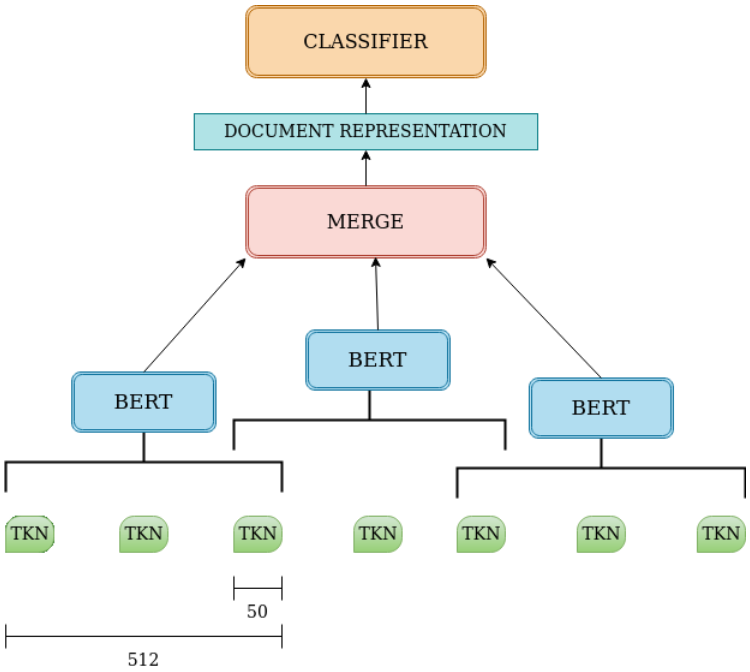


Figure 1. The document representation approach using a sliding window over the whole input sequence. Each subsequence is embedded using a fine-tuned BERT model, and all the subsequences are then merged into a final document representation, which is sent further as the input to the classifier. The length of the sliding window is 512 tokens. The first 50 tokens of each subsequent sliding window overlap with the last 50 tokens of the previous sliding window.

For each document x_i in the training set $S = \{x_1, x_2, \ldots, x_n\}$, we produced a document representation $d \in R^{1 \times t}$, where t is the dimension of the representation, by encoding the document with the mBERT model and taking the representation of the [CLS] token from the last layer. We sent this representation through a linear layer and a softmax function to map it to one of the predefined classes $C = \{y_1, y_2, \ldots, y_n\}$.

$$h = Linear(d, W) \quad (1)$$

$$\hat{y} = Softmax(h) \quad (2)$$

We calculated the loss of the sentiment classification task: $\mathcal{L}_s$ at the end using the negative log likelihood loss function

$$\mathcal{L}_s = -\log(\hat{y}_i)$$

where $\hat{y}_i$ is the probability of the correct class.
The final loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_s$$

where $\mathcal{L}_{mlm}$ represents the loss from the masked language modeling task. The model is then jointly trained on both tasks by backpropagating the final loss through the whole network.

The original mBERT model is pre-trained on another task, namely next sentence prediction, which, according to the authors, helps the model learn sentence relationships. During training, the input for this task is treated as belonging to two separate sequences and the model has to decide if the two sequences follow one another in the original text or not. This information is useful for a variety of downstream tasks such as question answering. Since in this experiment we are dealing with a classification task, where the input is treated as being a part of the same sequence, we felt the additional training using the next sentence prediction task would not add much relevant information to the model so we omitted it in the intermediate training phase.

5. Experimental Setup

This sections describes the setup that we used to perform the experiments. It is divided into three subsections: the first subsection describes the regime we used for the fine-tuning phases; the second subsection describes the regime we used for the intermediate training phase; and the third subsection presents the evaluation of the trained models.

5.1. Fine-Tuning Phase

For the fine-tuning phase, we used the Slovenian news dataset [3] annotated on the document level (see Section 3), as the goal of our classification is to assign the sentiment label to a news article. We followed the suggestions in the original paper by Devlin et al. [10] for fine-tuning. We used the Adam optimizer with the learning rate of $2E-5$ and learning rate warmup over the first 10% of the training instances. For regularization purposes, we used the weight decay set to 0.01. We reduced the batch size from 32 to 16 due to the high memory consumption during training, which was the result of a long sequence length. For benchmarking purposes, we used the k-fold cross-validation training regime for the fine-tuning phase, where we split the dataset into k folds. In each cross-validation step, the k-1 folds are used as the training set, while the k-th fold is used as the testing set. The models in each cross-validation step were trained for 3 epochs. To avoid overfitting, we split the training folds into smaller training and development sets. After each epoch, we measured the performance on the development set and saved the new model parameters only if the performance of the model on the development set increased. For the document representation methods, described in Sections 4.1 and 4.2, the fine-tuning of the language model and the training of the classification head were performed end-to-end, while for the methods, described in Section 4.3, the classification heads were trained after the fine-tuning phase was completed. Otherwise, the training regime and the chosen hyperparameters were the same for all the experiments.

5.2. Intermediate Training Phase Regime

For the intermediate training phase, we utilized the proposed modified modeling objectives, described in Section 4.4. We used the Slovenian news dataset with annotations on the paragraph level. The annotations on this level of granularity were used because we wanted to perform the intermediate training phase on a different dataset than the one used for fine-tuning, but containing information relevant for the document-level sentiment classification task.

Since the annotated paragraphs were part of the same documents we used for the fine-tuning step, we took measures to prevent any form of data leakage. As described in Section 5.1, the fine-tuning phase was performed using 10-fold cross-validation. We performed the intermediate training in each cross-validation step, but excluded the paragraphs that were part of the documents in the k-th testing fold of the fine-tuning step from the dataset. We split the remaining data into a training and development set and trained the language model for a maximum of five epochs. At the end of each epoch, we calculated the perplexity score of the model on the development set and saved the new weights only if perplexity improved in the previous epoch. If perplexity did not improve for three

consecutive epochs, we stopped the training early. For this phase, we used the same hyperparameter settings as for the fine-tuning phase.

5.3. Evaluation

All the models were first trained and evaluated on the Slovenian dataset using 10-fold cross-validation as described in Sections 5.1 and 5.2. Next, the performance of the models from each fold was additionally tested on the Croatian test set to check the performance in the zero-shot learning setting (i.e., without any Croatian data used in training). The performances from each fold on the Croatian test set were then averaged and reported as a final result. The results for this set of experiments are presented in Table 2. The performance of the models was summarized using a standard classification metric, namely the macro-averaged F1 score, which is the appropriate measure given the highly imbalanced nature of the dataset (dominant neutral class). For completeness, we also separately report the precision and recall, both macro-averaged over all classes. Additionally, we also report the average F1 score performance of the model on the Slovenian and Croatian test sets. The performance of our models was compared to the baseline majority classifiers for both the Slovenian and Croatian datasets.

Table 2. Results of the document representation approaches. The first column shows the performance of models in the Slovenian 10-fold cross-validation setting; the second column is the average zero-shot performance on the Croatian test set; and the last column presents the average F1 score of the results on the Slovenian and Croatian datasets. Best results are marked in bold.

Model	Slovenian Cross-Validation			Croatian Test Set			Average
	Precision	Recall	F1	Precision	Recall	F1	F1
Majority classifier	17.34	33.33	22.76	0.20	0.33	25.00	/
Beginning of the document	65.45 ± 2.61	**62.83** ± 2.46	63.34 ± 2.29	57.74 ± 1.20	**53.91** ± 2.41	52.06 ± 2.64	57.70
Beginning and end of the document	64.72 ± 2.82	62.67 ± 2.69	63.33 ± 2.56	**59.00** ± 1.62	53.53 ± 3.64	**52.41** ± 2.58	**57.87**
Sequences from every part of the document							
Most informative subsequence	64.42 ± 2.44	62.09 ± 2.27	63.00 ± 2.34	57.87 ± 1.32	53.23 ± 2.82	52.30 ± 2.86	57.65
Averaging subsequence representations	**66.50** ± 3.13	62.00 ± 2.45	**63.39** ± 2.42	57.53 ± 1.14	52.95 ± 3.38	51.55 ± 3.93	57.47
1D CNN	63.96 ± 10.02	60.91 ± 5.22	61.58 ± 7.78	54.96 ± 5.48	53.31 ± 3.62	50.28 ± 4.65	55.93

6. Results

This section presents the results of the experiments conducted in the course of this study. We first present the results of the document representation approaches. The results are presented in Table 2. Next, for the best performing representation approach, we test our newly introduced technique for sentiment classification with intermediate training, and the results with and without the intermediate training objective are compared in Table 3. We also compare our results with the previous sate-of-the-art SVM and Naive Bayes models on the Slovenian dataset from [3], as well as with the neural network model based on LSTMs and TF-IDF from [39]. We note, however, that the testing regime in these experiments was not the same. In [3], the authors tested their models using five times 10-fold cross-validation, while in [39], the model was trained and tested on a random train-test split of the whole dataset with an 80:20 train-test split ratio. For this reason, the results are not directly comparable.

Table 3. Performance of the model using our intermediate sentiment classification training approach compared to the model without intermediate training. Additionally, we include the reported results from the related work using the same dataset. Best results are marked in bold.

Model	Slovenian			Croatian			Average
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1**
Majority classifier	17.34	33.33	22.76	0.20	0.33	25.00	/
Reported results from related studies							
SVM (from Bučar et al. [3]) 5 × 10 CV	/	/	63.42 ± 1.96	/	/	/	/
NBM(from Bučar et al. [3]) 5 × 10 CV	/	/	65.97 ± 1.70	/	/	/	/
LSTM+TF-IDF (from Pelicon [39])train-set split	/	/	62.5	/	/	/	/
Results from the current study							
Beginning of the document	65.45 ± 2.61	62.83 ± 2.46	63.34 ± 2.29	**57.74** ± 1.20	53.91 ± 2.41	52.06 ± 2.64	57.70
Beginning and end of the document with sentiment intermediate training	**67.19** ± 2.67	**66.00** ± 3.00	**66.33** ± 2.60	56.32 ± 1.88	**54.90** ± 2.36	**54.77** ± 1.39	**60.55**

As shown in Table 2, all the models using one of the tested document representation methods in this experiment performed better than the majority baseline classifier by a substantial margin. The best performing model on the Slovenian dataset (in terms of F1 score) utilizes document representations formed by simple averaging of the subsequence representations. The different document representation methods that were tested in this work do not seem to impact the model performance much as the performances of all our models differed only by a small margin when tested on the Slovenian data.

As far as absolute performance, we can see that the tested methods achieved F1 scores in the sixties for this particular Slovenian dataset with the best F1 score of 63.39 with averaging subsequence representations. When these models were tested on the Croatian test set in a zero-shot setting, the performance additionally dropped for approximately 11% with best the F1 scores achieving the low fifties. The best performing representation on the Croatian dataset uses the beginning and end of the document. Interestingly, the best performing model on the Slovenian dataset also saw the highest drop on the Croatian dataset of 11.84%. We additionally observed high variance of the CNN model compared to the other models.

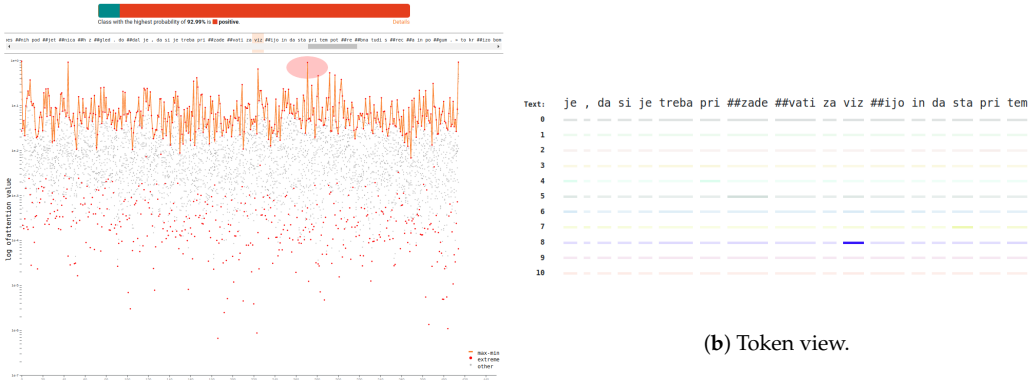
Since the three best performing document representation techniques were within a 0.06% difference on the Slovenian dataset, for experiments with intermediate training for sentiment enrichment, we opted for the document representation that used the beginning and ending of the input document as its average performance on the test sets of both Slovenian and Croatian languages was the highest. The results for the intermediate training experiment (Table 3) show that the model with the additional intermediate training step outperforms the model without the intermediate training step when using the same document representation technique. The results show three points better average performance on the Slovenian dataset and 2.68 points average improvement on the Croatian dataset in terms of the F1 score. Our model also manages to outperform the previous state-of-the-art models on the Slovenian dataset, achieving a 0.36 point increase in terms of F1 score, however this should be taken with precaution as the two evaluation settings differ.

7. Qualitative Exploration of the Models: Behavior of the Attention Space

With the increasing use of neural language models, in recent years, the methodology aimed at the exploration of the human-understandable patterns, emerging from trained models, has gained notable attention. Models, such as BERT [40] and similar ones, can consist of hundreds of millions of parameters, which carry little useful information in terms of studying which parts of the model input were of relevance when making a prediction. To remedy this shortcoming, visualization methodologies are actively developed and researched for the task of better understanding the associations between the input token space and the constructed predictions.

The existing toolkits that offer the exploration of attention have been actively developed in recent years [41,42] and are widely used to better understand a given model's behavior. In this section, we exploit the recently introduced, freely available AttViz [43], an online toolkit for the exploration of the self-attention space of trained classifiers (http://attviz.ijs.si/). The tool is used to explore the behavior of the self-attention when considering positive, negative, and neutral classifications. The original tool was developed for instance-based exploration. In addition, we introduce a novel functionality of the tool aimed at the analysis of global attention values (per class analysis on the token collection level).

In the remainder of this section, we fist present a collection of selected examples, offering insight into the trained model's behavior. We begin by discussing selected positive instances, followed by neutral and negative ones. All the visualizations were done with the sentiment-enriched model that we trained in the course of this study. The main aim of this section is to explore the currently available means of inspecting trained neural language models. A positive example is shown in Figure 2.



(**a**) Sequence view.

(**b**) Token view.

Figure 2. Positive Example No. 41. The red ellipse (**a**) highlights one of the tokens (byte pairs) with the highest (normalized) self-attention—the token is part of the word "vizija" (translation: vision) (**b**). Note also the peaks at the beginning and the end; these peaks refer to the special tokens (e.g., [CLS]]).

The positive example was selected as it has a very high probability of being positive class and it showcases two main patterns that can be observed throughout the space of positively classified examples: first, only a handful of tokens are emphasized (if any), and second, there appears to be strong bias towards the first and the last token, indicating the potential effect of pre-training.

Next, we considered some of the examples classified as negative sentiment (see the example in Figures 3 and 4).

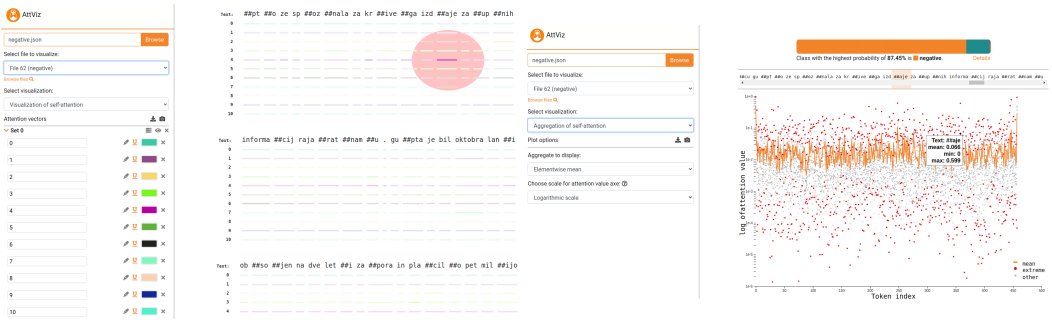


Figure 3. Negative Example No. 62. In this example, one of the highest attention values was around the token "izdaje" (translation: treason), which could be one of the carriers of the negative sentiment. Note that individual lines represent attention values for each of the ten attention heads. The document was classified with 87.45% probability.

The attention (highlighted red circle) peaks at the discussed token (translated as treason and negotiations respectively) can be observed, indicating that the neural language model picked up a signal at the token level during the association of the byte-paired inputs with outputs. Furthermore, we observed a similar pattern related to the starting [CLS] token, as well as the ending [SEP] token, i.e., token defining the end of the sentence. The pattern was consistent also throughout the neutral examples.

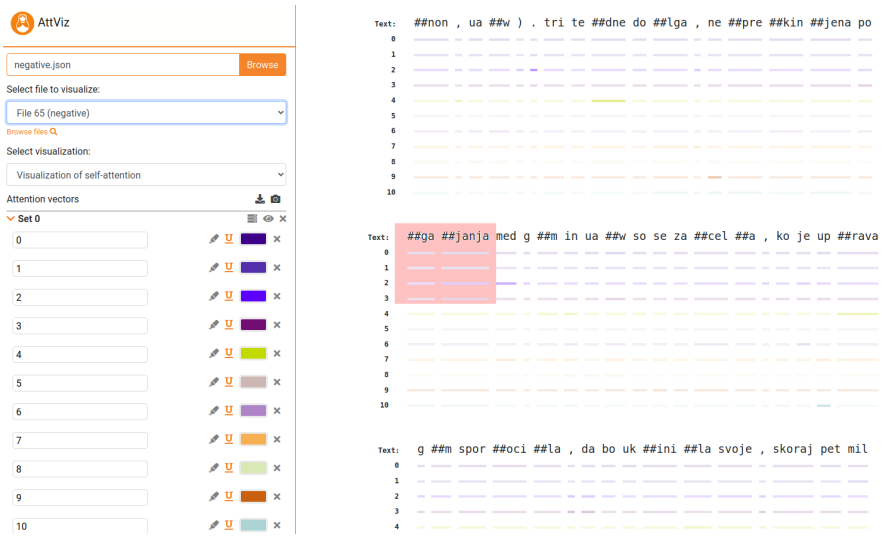


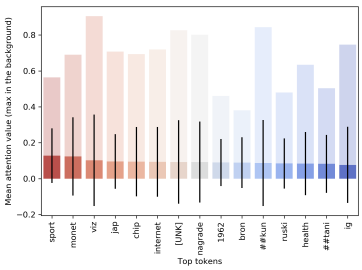
Figure 4. Negative Example No. 65. The highlighted region (red) corresponds to the term "pogajanja" (translation: negotiations), which appears to be associated with the classification of the observed text into the negative class.

The considered attention spaces offered insight into two main aspects of the trained model. First, the self-attention space, i.e., the space of the attention values alongside the attention matrix diagonals, offers relatively little insight into what the model learned. There are at least two main reasons for the observed behavior, as it appears to deviate from the reported explanations [43]. First, the considered documents are relatively long. Such documents give rise to a higher spread of the self-attention, smoothing out the individual peaks. Second, the wider spread of the attention could also be to the morphology-rich language considered (Slovene).

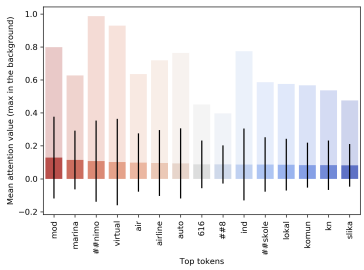
We next discuss the behavior of the global attention values both at the token, as well as the distribution level. The top 15 tokens according to the mean attention values are shown in Figure 5.

The presented results confirm the initial finding (e.g., Figure 2) that most of the attention space has high variability and, as such, does not directly offer interpretable insights; however, some meaningful results are also observed, e.g., the token with the greatest attention value for the positive class is sport. The final analysis we conducted was at the level of the global attention distributions. Here, we plotted the kernel density estimates of raw attention values across different types of instances. The results are shown in Figure 6.

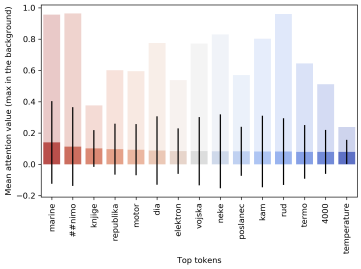
The distribution visualization indicates that the main differences emerge when considering the minimum value, a given token ever achieved; this result, albeit unexpected, potentially indicates that the attention is for classification of negative texts focused on a more particular subset of tokens, yielding a lower average subject to a skewed distribution. We finally offer quantile-quantile plots in Figure 7.



(**a**) Positive sentiment.

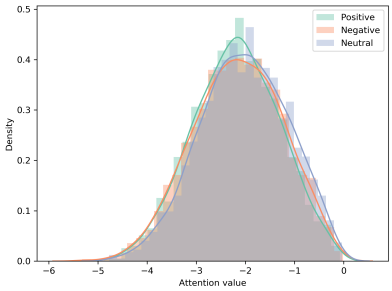


(**b**) Negative sentiment.

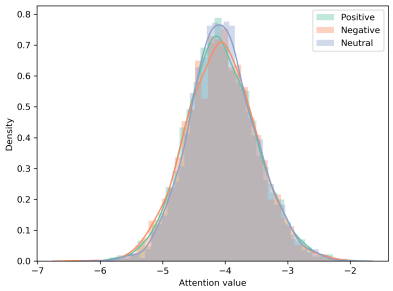


(**c**) Neutral sentiment.

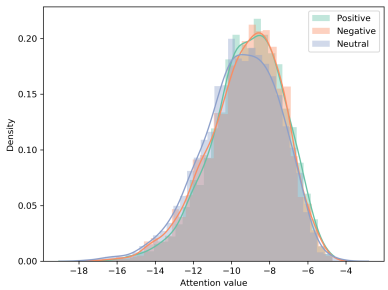
Figure 5. Visualization of token level attention. The figures represent the top 15 tokens according to the mean attention values. In the background, the maximum attention for a given token is also plotted. Note that the high standard deviation indicates little emphasis on the individual tokens.



(**a**) Maximum attention per token.

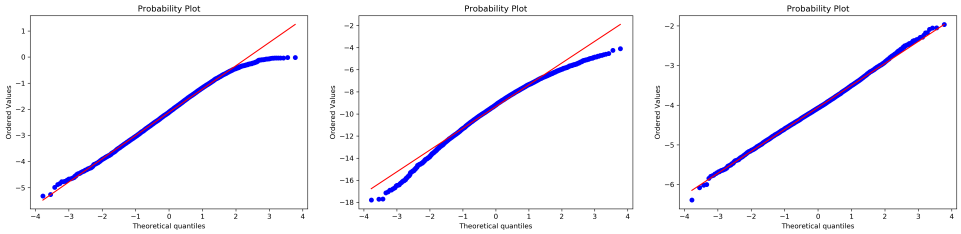


(**b**) Mean attention per token.



(**c**) Minimum attention per token.

Figure 6. Visualization of attention (log-transformed) distributions. It can be observed that the largest differences emerge when considering minimum attention. There, the negative texts' distribution is the most skewed. When considering maximum and mean distributions, however, no notable differences emerge.



(**a**) QQ plot: max attention. (**b**) QQ plot: min attention. (**c**) QQ plot: mean attention.

Figure 7. The quantile-quantile fits of the three considered attention distributions. It can be observed that the min and max attention distributions are skewed, indicating the presence of more extreme values.

The considered QQ-plots further confirm the observation that the skewed distribution of attention can be observed when considering min-max values; however, on average, the log transform could be interpreted to behave as a normal distribution; however, additional tests, such as Pearson's sample skewness (computed as $\frac{n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}}$, where x_i is the i-th value out of n samples) could be conducted to further quantify the attention behavior.

8. Availability

The croatian news dataset with document-level sentiment annotations is available on the CLARIN repository under the Creative Commons license (CC-BY-NC-ND) (http://hdl.handle.net/11356/1342). The code for all the experiments is available on GitHub (https://github.com/PeliconA/crosslingual_news_sentiment.git).

9. Conclusions and Future Work

In this work, we addressed the task of sentiment analysis in news articles performed in a zero-shot cross-lingual setting. The goal was to successfully train models that could, when trained on data in one language, perform adequately also on data in another language. For this purpose, we used publicly available data of Slovenian news manually labeled for sentiment to train our models. Additionally, we gathered a new dataset of Croatian news and labeled it according to the guidelines for the annotation of the Slovenian dataset. This new dataset served as a test set for the zero-shot cross-lingual performance of our models.

We based our models on the multilingual Transformer-based model BERT, which has shown remarkable multilingual and cross-lingual performance. We however identified two potential drawbacks with the BERT model. The input window of the BERT model is fixed and relatively short. A widespread approach to this limitation is to shorten the input before sending it to the model for processing. While this approach is adequate for shorter texts, with longer documents, like news articles, it may cause severe information loss. The second drawback is that while BERT is pre-trained on a large collection of data, the only explicit sentiment signal it gets is during the fine-tuning phase on a usually small collection of labeled data.

To remedy the first potential drawback, we first tested several techniques for producing more informative long document representations. The techniques, which were described in detail, were partially inspired by earlier work, but to the best of our knowledge, they have not yet been tested in a cross-lingual setting. Our results show that all the techniques outperform the majority baseline classifier by a large margin, even when applied to the Croatian test set in a zero-shot setting where the model is not fine-tuned on Croatian data.

For the second identified limitation of the BERT model, we proposed a novel intermediate learning phase that encompasses the masked language modeling task and sentiment classification task. This phase is performed before the fine-tuning phase using a training set with separate annotations. The goal of this

phase is to induce the sentiment-related information directly into the BERT representations before the fine-tuning begins on the target task data. Results show that after fine-tuning, the sentiment-enriched model outperforms the models without the intermediate training phase both on the Slovenian dataset and on the Croatian test set in a zero-shot setting. Additionally, it slightly outperforms the current state-of-the-art on the Slovenian dataset, as reported in [3].

In the future, we plan to further test our proposed intermediate sentiment-enrichment phase with masked language modeling and sentiment classification tasks. Currently, the fine-tuning and the intermediate training phases share the dataset, but use labels on different levels of granularity: we used document-level labels for fine-tuning and paragraph-level labels for intermediate training. We would like to test how using training data from a very different training set would impact the performance of the proposed intermediate training step. We will also test the general transferability of this phase. Given a large enough corpus of sentiment-labeled instances that can be used for the intermediate training step, we would like to see if a Transformer-based model enriched with our proposed method can work well on sentiment tasks in different target languages and from different domains. Another interesting research area would be using topic modeling as a supplementary method for the news-related sentiment classification task. Such research would also test the underlying assumption that there is a positive correlation between the topic of a news article and the sentiment that a news article evokes in the readers. Even though the news articles in the datasets used for this work are not explicitly labeled for topics, they nevertheless deal with varying content and could support such research.

Author Contributions: A.P. and S.P. designed the study and developed its methodology. M.P. and D.M. provided the data for the study and guided the annotation process. Formal analysis of the study was done by A.P. Software for the experiments was written by A.P. Visualization of the trained models was done by B.Š. Validation of the results and supervision of the study was done by S.P. A.P., M.P., D.M., B.Š. and S.P. cotributed to the writing, reviewing and editing of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The work of A.P. was funded also by the European Union's Rights, Equality and Citizenship Programme (2014–2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, Grant No. 875263). We acknowledge also the funding by the Slovenian Research Agency (ARRS) core research program Knowledge Technologies (P2-0103). The results of this publication reflect only the authors' views, and the Commission is not responsible for any use that may be made of the information it contains.

Acknowledgments: We would like to thank 24sata, especially Hrvoje Dorešić and Boris Trupčević, for making the data available. We thank Jože Bučar for leading the annotation process.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Details on Croatian Dataset Construction

For the selection of articles, the time period was specified: approximately half of the articles were selected from the period from 1 September 2007 to 31 December 2013 in order to match the Slovenian dataset, while the other half were recent articles from last five years. From the initial set of articles, short and medium length articles were kept, leading to final selection. The articles were then cleaned and preprocessed, and the quality was checked (automatically and manually). The final dataset consisted of 2025 news articles. The sentiment annotation task was performed on three levels: document, paragraph, and sentence level.

For the selection of annotators, the condition of being native speakers was imposed, and we also considered the candidate's interest in the task.

The annotator were trained in two phases:

- In the first phase, we introduced the project EMBEDDIA and its goals. A referee introduced the web application for the annotation task. The annotators received basic guidelines, which were explained to them in detail by a referee. This was followed by the annotation of five articles,

which were annotated together on the three levels (sentence, paragraph, and document level). Using a five-level Likert scale: [35] (1—very negative, 2—negative, 3—neutral, 4—positive, and 5—very positive), the annotators annotated each article according to the following question: "Did this news evoke very positive/positive/neutral/negative/very negative feelings? (Please specify the sentiment from the perspective of an average Croatian web user)". Together with a referee, they discussed the individual instances, every single decision, and the annotation grade and resolved possible issues and doubts.

- In the second phase, all annotators annotated the same 25 articles individually. Afterwards, we analyzed the results of the annotation. The agreement (Cronbach's alpha measure) between the annotators on the document level was 0.816, which was a very good achievement with only 25 articles. We planned to achieve a 0.8 threshold. If the annotators had not achieved the planned threshold, they would repeat the second phase until they achieved it. The instances with lower agreement were discussed, and the issues were resolved.

Since a satisfying inter-annotator agreement was reached, the rest of the 2000 were annotated by different numbers of annotators. They followed the instructions they were given in the first and second phases.

To evaluate the process of annotation, we explored correlation coefficients using various measures of inter-annotator agreement at three levels of granularity, as shown in Table A1. The first three internal consistency estimates of reliability for the scores, shown in Table A1, normally range between zero and one. The values closer to one indicate more agreement, when compared to the values closer to zero. Cronbach's alpha values indicated a very good internal consistency at all levels of granularity. Normally, we refer to a value greater than 0.8 as a good internal consistency and above 0.9 as an excellent one [44]. The value of Krippendorff's alpha [45] at the document level of granularity implied a fair reliability test, whereas its values at the paragraph level and sentence level were lower. Fleiss' kappa values illustrated a moderate agreement among the annotators at all levels of granularity. In general, a value between 0.41 and 0.60 implies a moderate agreement, above 0.61 a substantial agreement, and above 0.81 an almost perfect agreement [46]. Kendall's values indicated a fair level of agreement between the annotators at all levels of granularity. Correspondingly, the Pearson and Spearman values range from –1 to 1, where 1 refers to the total positive correlation, 0 to no correlation, and –1 to the total negative correlation. The coefficients showed moderate positive agreement among the annotators, but their values decreased when applied to the paragraph and the sentence level. Usually, the values above 0.3 refer to a weak correlation, above 0.5 to a moderate correlation, and above 0.7 to a strong correlation [47].

Table A1. Results of dataset annotation: level of inter-rater agreement for document, paragraph, and sentence levels.

	Document Level			**Paragraph Level**			**Sentence Level**		
a_c	0.927			0.888			0.881		
a_k	0.671			0.565			0.548		
k	0.527			0.489			0.441		
	min	*max*	*avg*	*min*	*max*	*avg*	*min*	*max*	*avg*
r_p	0.544	0.824	0.682	0.488	0.719	0.572	0.425	0.706	0.558
r_s	0.557	0.762	0.669	0.474	0.702	0.548	0.42	0.696	0.54
W	0.508	0.73	0.625	0.449	0.656	0.513	0.389	0.649	0.504

Our results support the claim by [48] that it can be more difficult to accurately annotate sentences (or even phrases). In general, the sentiment scores by different annotators were more consistent at the document level than at the paragraph and sentence level.

The final sentiment of an instance is defined as the average of the sentiment scores given by the different annotators (as in the Slovenian news set). An instance was labeled as:

- negative, if the average of given scores was less than or equal to 2.4,
- neutral, if the average of given scores was between 2.4 and 3.6,
- positive, if the average of given scores was greater than or equal to 3.6.

References

1. Beigi, G.; Hu, X.; Maciejewski, R.; Liu, H. An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment Analysis and Ontology Engineering*; Springer: Cham, Switzerland, 2016; pp. 313–340.
2. Mejova, Y. *Sentiment Analysis: An Overview*; University of Iowa, Computer Science Department: Iowa City, IA, USA, 2009.
3. Bučar, J.; Žnidaršič, M.; Povh, J. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Lang. Resour. Eval.* **2018**, *52*, 895–919. [CrossRef]
4. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [CrossRef]
5. Van de Kauter, M.; Breesch, D.; Hoste, V. Fine-Grained Analysis of Explicit and Implicit Sentiment in Financial News Articles. *Expert Syst. Appl.* **2015**, *42*, 4999–5010. [CrossRef]
6. Bhutani, B.; Rastogi, N.; Sehgal, P.; Purwar, A. Fake news detection using sentiment analysis. In Proceedings of the IEEE 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2019; pp. 1–5.
7. El Ali, A.; Stratmann, T.C.; Park, S.; Schöning, J.; Heuten, W.; Boll, S.C. Measuring, understanding, and classifying news media sympathy on twitter after crisis events. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–13.
8. Rambaccussing, D.; Kwiatkowski, A. Forecasting with news sentiment: Evidence with UK newspapers. *Int. J. Forecast.* **2020**. [CrossRef]
9. Bowden, J.; Kwiatkowski, A.; Rambaccussing, D. Economy through a lens: Distortions of policy coverage in UK national newspapers. *J. Comp. Econ.* **2019**, *47*, 881–906. [CrossRef]
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Schwartz, B. Google's Latest Search Algorithm to Better Understand Natural Language. Search Engine Land. 25 October 2019. Available online: https://searchengineland.com/welcome-bert-google-artificial-intelligence-for-understanding-search-queries-323976 (accessed one 28 August 2020).
12. Albarino, S. Does Google's BERT Matter in Machine Translation? Slator. 17 October 2019. Available online: https://slator.com/machine-translation/does-googles-bert-matter-in-machine-translation/ (accessed one 28 August 2020).
13. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
14. Karthikeyan, K.; Wang, Z.; Mayhew, S.; Roth, D. Cross-lingual ability of multilingual bert: An empirical study. In Proceedings of the International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2019.
15. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Brew, J. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
16. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised data augmentation for consistency training. *arXiv* **2019**, arXiv:1904.12848.
17. Lin, K.Y.; Yang, C.; Chen, H.H. Emotion Classification of Online News Articles from the Reader's Perspective. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9–12 December 2009; Volume 1, pp. 220–226. [CrossRef]
18. Li, X.; Xie, H.; Chen, L.; Wang, J.; Deng, X. News Impact on Stock Price Return via Sentiment Analysis. *Knowl. Based Syst.* **2014**, *69*. [CrossRef]

19. Mansar, Y.; Gatti, L.; Ferradans, S.; Guerini, M.; Staiano, J. Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment towards Brands from Financial News Headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 817–822. [CrossRef]
20. Moore, A.; Rayson, P. Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 581–585. [CrossRef]
21. Wan, X. Co-training for cross-lingual sentiment classification. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 235–243.
22. Guzmán, F.; Chen, P.J.; Ott, M.; Pino, J.; Lample, G.; Koehn, P.; Chaudhary, V.; Ranzato, M. The FLoRes evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv* **2019**, arXiv:1902.01382.
23. Zhou, X.; Wan, X.; Xiao, J. Attention-based LSTM network for cross-lingual sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 247–256.
24. Chen, Z.; Shen, S.; Hu, Z.; Lu, X.; Mei, Q.; Liu, X. Emoji-powered representation learning for cross-lingual sentiment classification. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 251–262.
25. Funaki, R.; Nakayama, H. Image-mediated learning for zero-shot cross-lingual document retrieval. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 585–590.
26. Rei, M.; Søgaard, A. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. *arXiv* **2018**, arXiv:1805.02214.
27. Wang, Y.; Che, W.; Guo, J.; Liu, Y.; Liu, T. Cross-lingual BERT transformation for zero-shot dependency parsing. *arXiv* **2019**, arXiv:1909.06775
28. Hsu, T.Y.; Liu, C.L.; Lee, H.Y. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. *arXiv* **2019**, arXiv:1909.09587
29. Ye, Z.; Geng, Y.; Chen, J.; Chen, J.; Xu, X.; Zheng, S.; Wang, F.; Zhang, J.; Chen, H. Zero-shot Text Classification via Reinforced Self-training. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 3014–3024.
30. Jebbara, S.; Cimiano, P. Zero-Shot Cross-Lingual Opinion Target Extraction. *arXiv* **2019**, arXiv:1904.09122
31. Fei, H.; Li, P. Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 5759–5771.
32. Tan, X.; Yan, R.; Tao, C.; Wu, M. Classification over Clustering: Augmenting Text Representation with Clusters Helps! In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*; Springer: Cham, Switzerland, 2019; pp. 28–40.
33. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical Transformers for Long Document Classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 838–844.
34. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
35. Likert, R. A Technique for the Measurement of Attitudes. *Arch. Psychol.* **1932**, *140*, 5–55.
36. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144
37. Pruksachatkun, Y.; Phang, J.; Liu, H.; Htut, P.M.; Zhang, X.; Pang, R.Y.; Vania, C.; Kann, K.; Bowman, S.R. Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? *arXiv* **2020**, arXiv:2005.00628.

38. He, C.; Chen, S.; Huang, S.; Zhang, J.; Song, X. Using Convolutional Neural Network with BERT for Intent Determination. In Proceedings of the IEEE 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 65–70.
39. Pelicon, A. Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami. In Proceedings of the Conference on Language Technologies and Digital Humanities 2020 (to appear), Ljubljana, Slovenia, 17–20 March 2020.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
41. Vig, J. Visualizing Attention in Transformer-Based Language Representation Models. *arXiv* **2019**, arXiv:1904.02679.
42. Vig, J.; Belinkov, Y. Analyzing the structure of attention in a transformer language model. *arXiv* **2019**, arXiv:1906.04284.
43. Škrlj, B.; Eržen, N.; Sheehan, S.; Luz, S.; Robnik-Šikonja, M.; Pollak, S. AttViz: Online exploration of self-attention for transparent neural language modeling. *arXiv* **2020**, arXiv:2005.05716.
44. George, D.; Mallery, P. *SPSS for Windows Step-by-Step: A Simple Guide and Reference, 14.0 Update*, 7th ed.; Allyn and Bacon, Inc.: Boston, MA, USA, 2006.
45. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 2nd ed.; Sage Publications: Thousand Oaks, CA, USA, 2004.
46. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
47. Rumsey, D.J.; Unger, D. *U Can: Statistics for Dummies*; John Wiley: Hoboken, NJ, USA, 2015.
48. O'Hare, N.; Davy, M.; Bermingham, A.; Ferguson, P.; Sheridan, P.; Gurrin, C.; Smeaton, A. Topic-dependent sentiment analysis of financial blogs. In Proceedings of the TSA 2009—1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong, China, 6 November 2009; TSA: Arlington County, VA, USA, 2009; pp. 9–16, ISBN 978-1-60558-805-6.

Appendix D: COVID-19 v slovenskih spletnih medijih: analiza s pomočjo računalniške obdelave jezika

PETINPETDESETLET
slovensko sociološko društvo

PANDEMIČNA DRUŽBA

SLOVENSKO SOCIOLOŠKO SREČANJE
Ljubljana, 24.–25. september 2021

Ljubljana, 2021

Izdajatelj:
Slovensko sociološko društvo
Kardeljeva ploščad 5, 1000 Ljubljana

Uredniki:
Miroljub Ignjatović, Aleksandra Kanjuo Mrčela, Roman Kuhar

Tehnični urednik:
Igor Jurekovič

Programski odbor:
Predsedstvo Slovenskega sociološkega društva

Recenzentke:
Anja Zalta, Alenka Švab in Veronika Tašner

Oblikovanje in prelom:
Polonca Mesec Kurdija

Korekture:
avtorji

Elektronska izdaja:
Publikacija je brezplačno dostopna na elektronskem naslovu:
http://www.sociolosko-drustvo.si/.

Ljubljana, 2021

Kataložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 77660931

ISBN 978-961-94302-6-2 (PDF)

SENJA POLLAK
Inštitut Jožef Stefan
MATEJ MARTINC
Inštitut Jožef Stefan
ANDRAŽ PELICON
Inštitut Jožef Stefan
MATEJ ULČAR
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
ANDREJA VEZOVNIK
Univerza v Ljubljani, Fakulteta za družbene vede

COVID-19 V SLOVENSKIH SPLETNIH MEDIJIH: ANALIZA S POMOČJO RAČUNALNIŠKE OBDELAVE JEZIKA

Povzetek: *V prispevku s pomočjo metod za računalniško obdelavo naravnega jezika analiziramo poročanje slovenskih medijev o epidemiji covid-19. V prispevku najprej identificiramo osrednje teme poročanja na petdesetih slovenskih novičarskih portalih v obdobju 1.1.2020–31.12.2020. Nato z uporabo kontekstualnih besednih vložitev analiziramo razlike v poročanju na manjšem korpusu štirinajstih izbranih novičarskih portalov. Ugotavljamo, da se razlike med portali kažejo predvsem pri besedah povezanih s spornimi temami v javnosti. Na primeru besede kolesar, pokažemo, da se na portalih nova24TV.si, demokracija.si in necenzurirano.si, beseda kolesar intenzivneje povezuje z besedo covid-19, kot na primer na portalu siol.net, ter da beseda na prvih treh portalih nosi politično konotacijo, medtem ko je na siol.net povezana predvsem s športom.*

Ključne besede: *covid-19, novičarski mediji, obdelava naravnega jezika, besedne vložitve, računalniška analiza besedil.*

Uvod

Epidemija covid-19 je v zadnjem letu sprožila širšo družbeno krizo. V času upravljanja s krizo igrajo pomembno vlogo novičarski spletni mediji, saj le-ti služijo javnosti kot primarni vir pridobivanja informacij. Spletni mediji so namreč v Sloveniji v zadnjih letih postali prevladujoči vir za spremljanje novic (iPROM in Valicon 2021). Pomen proučevanja medijskih vsebin ima dolgo zgodovino v komunikologiji. McCombs in Shaw (1972) sta denimo preučevala vpliv televizijske agende na neodločene volivce. Gerbner in dr. so raziskovali, kako televizijske vsebine vzgajajo medijsko občinstvo. Entman (1994) je preučeval, kako mediji na specifične načine okvirjajo teme in kako izbori in poudarki vsebin, vplivajo na dojemanje vsebin s strani občinstva. Cantril (1999) je pokazal, da so ljudje v času družbenih kriz posebej sugestibilni za medijske vsebine. Tradicija ukvarjanja s povezavo med medijskimi vsebinami in občinstvi je teoretsko kompleksna in nam v tem prispevku služi kot iztočnica za utemeljitev, da je preučevanje medijskih vsebin, posebej novičarskih, že več kot stoletje predmet osrednjega komunikološkega zanimanja.

Vrsta novejših študij proučuje medijske vsebine v povezavi s covid-19. S pomočjo modeliranja tematik Liu in dr. (2020) ugotavljajo pomen medijskega poročanja o covid-19 na Kitajskem. Rebello in dr. (2020) preučujejo, kako so se novičarske vsebine povezane s covid-19 manifestirale na spletnih družbenih omrežjih. Mutua in Ong'ong'a (2020) sta s pomočjo analize vsebin in okvirjanja analizirala poročanje mednarodnih tiskovnih agencij o covid-19. Podobno se s pomočjo analize okvirov Hubner (2021) loti novičarskih virov v ZDA. Hart in dr. (2020) s pomočjo računalniško podprte analize ugotavljajo stopnjo politizacije in polarizacije novic o covid-19 v novičarskih medijih v ZDA. Metodološko je najbolj sorodna diahrona analiza poročanja o covid-19 z uporabo gručenja pomenov s kontekstualnimi vložitvami (Montariol in dr. 2021). Zaenkrat so študije novičarskih vsebin v povezavi s covid-19 še vedno maloštevilne. V Sloveniji pa take študije še nimamo.

Naš korpus zajema več deset tisoč člankov. Ročna analiza velikih podatkov je časovno neizvedljiva. V prispevku pokažemo, kako lahko z metodami obdelave naravnega jezika ponudimo nov vpogled v poročanje slovenskih novičarskih portalov o epidemiji. Najprej uporabimo metodo modeliranja tematik s pomočjo latentne Dirichletove alokacije, v nadaljevanju pa metode, ki temeljijo na besednih vektorskih vložitvah. Besedne vektorske vložitve so večstodimenzionalne vektorske predstavitve besed, ki opisujejo besede glede na besedilni kontekst, v katerem se pojavljajo, natrenirane pa so z uporabo nevronskih mrež. Bližina med vektorji v vektorskem prostoru pa odraža semantično povezanost besed. Pri statičnih vložitvah eni besedi ustreza en vektor, pri kontekstualnih pa vektor predstavalja posamezno besedno rabo. Z metodami, ki temeljijo na kontekstualnih besednih vložitvah lahko tako tudi primerjamo rabe besed v različnih medijih.

Korpus

V pričujočem članku obravnavamo korpus o covid-19, ki zajema 89.204 člankov iz obdobja 1. 1.2020–31.12.2020. Z uporabo storitve EventRegistry (Leban in dr. 2014) smo zajeli članke, ki vsebujejo eno izmed besed *covid, koronavirus, sars-cov-2, covid19, covid-19, korona virus, koronavirusna, koronavirusen.* V korpus smo vključili članke tistih portalov, ki so zavedeni v Razvidu medijev, vodenega s strani Ministrstva za kulturo RS. Ta korpus člankov petdesetih portalov[1] (Korpus-50) smo nato uporabili za modeliranje tematik. Za analizo razlik med portali pa smo se omejili le na novičarske portale, ki so v korpusu imeli vsaj 1000 člankov, in izločili portale specializirane za športne in lokalne vsebine. Ta pod-izbor (Korpus-14) zajema članke iz rtvslo.si, siol.net, delo.si, žurnal24.si, vecer.com, 24ur.com, novice.svet24.si, reporter.si, dnevnik.si, demokracija.si, nova24tv.si, politikis.si, mladina.si in necenzurirano.si.

Modeliranje tematik

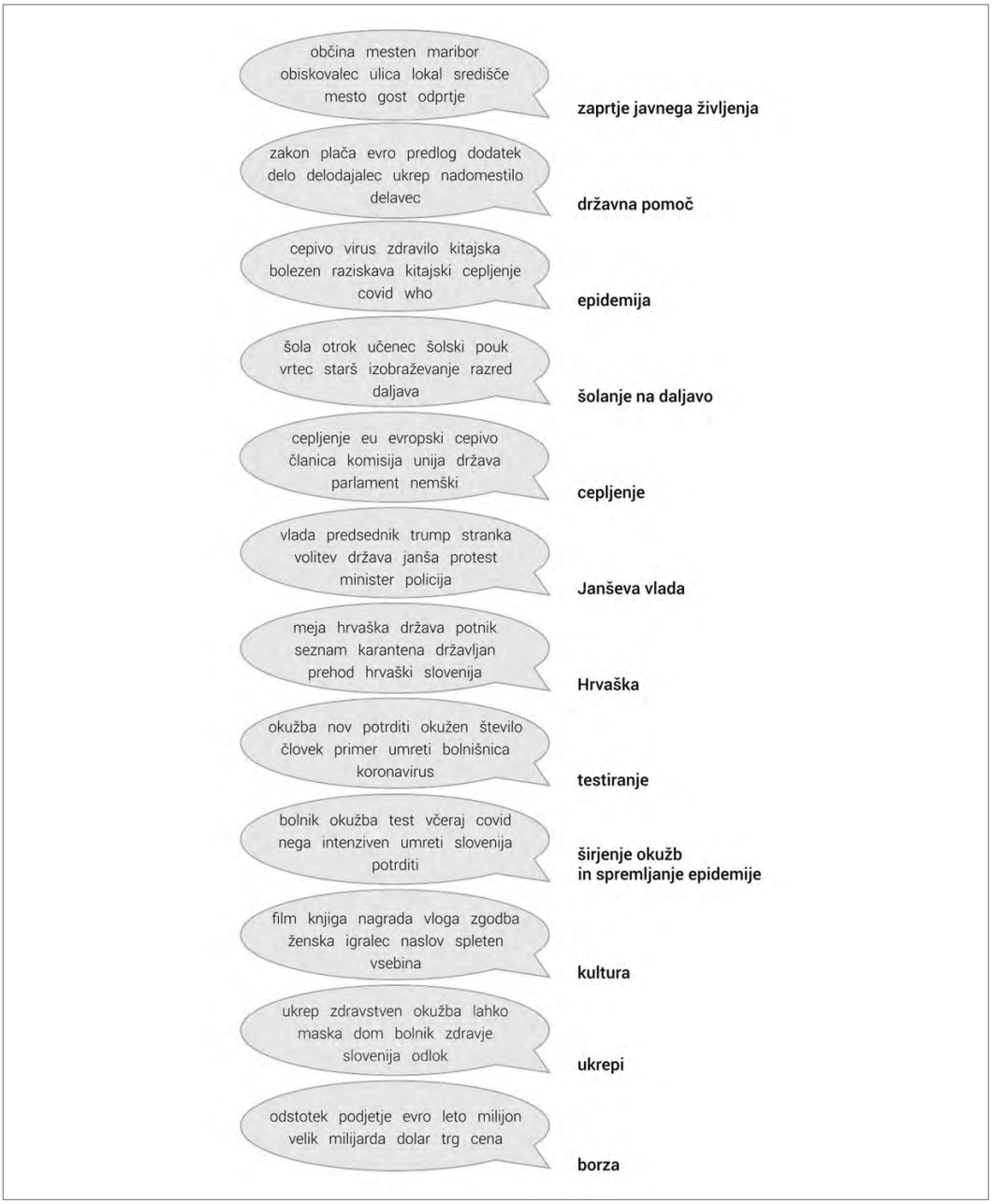
Z uporabo metode latentne Dirichletove alokacije - LDA (Blei in dr. 2003) smo avtomatsko prepoznali tematike v naboru Korpus-50. Metoda predvideva, da so dokumenti iz korpusa sestavljeni iz več tem, pri čemer je večja verjetnost, da vsak dokument obravnava manjše število tem. Podobno velja za besede: vsaka beseda z večjo verjetnostjo pripada manjšemu številu tem. Pred uporabo LDA smo besedila lematizirali, spremenili velike

1. https://docs.google.com/document/d/1gVpYwjCcmjuwVXDuNFKZp4FefXXZqFa7pACnsZnskg/edit?usp=sharing

začetnice v male, odstranili nepolnopomenske besede in utežili besede z mero TF-IDF, ki daje poudarek bolj specifičnim besedam za dokument.

Kot rezultat dobimo skupine besed, ki predstavljajo 20 najpogostejših tematik. Imena tematik smo določili ročno. Iz rezultatov smo odstranili teme, ki so vsebovale veliko šumnih podatkov ali se nam niso zdele zanimive za analizo (npr. športni dogodki). Končni seznam 12 tematik je prikazan na Sliki 1.

Slika 1: Najpogostejše tematike (metoda LDA).



Ekstrakcija besednih vložitev

Besedne vektorske vložitve, ki so natrenirane z uporabo nevronskih mrež, so predstavitve besed v prostoru, kjer vsako besedo opisuje vektor z več sto dimenzijami. Besede, ki so si blizu v vektorskem prostoru (kar lahko merimo s kosinusno razdaljo), so si tudi semantično podobne.

Pri statičnih vložitvah je posamezna beseda v korpusu predstavljena z enim vektorjem. Če reprezentacijo Korpusa-50 generiramo z modelom fastText[2] (Bojanowski in dr. 2017), so besedi koronavirus najbližje besede *nov, virus, enterovirus, pozitiven, testiranje, test, razširiti, izvid, okužba*. Za besedo *Janša* pa med 10 najbližjimi besedami najdemo tudi besedo *Šarec* in politike Višegrajske skupine ter desne evropske politike (Morawieck, Orban, Kurz).

Za razliko od statičnih vložitev, kjer vsako besedo predstavlja en vektor, pri kontekstualnih vložitvah vsako pojavitev besede opisuje svoj vektor. To je pomembno predvsem z vidika večpomenskih besed ali kjer analiziramo razlike med besedami v različnih kontekstih. Za eksperimente v nadaljevanju smo Korpus-14 lematizirali, kontekstualne vložitve pa smo zgradili z uporabo modela SloBERTA[3] (Ulčar in Robnik-Šikonja 2020). Povprečenje kontekstualnih reprezentacij na nivoju posamezne leme (osnovne oblike besede) (cf. Martinc in dr. 2020) v podkorpusu specifičnega medija nam omogoča primerjavo različnih medijskih vsebin.

Raznolikost poročanja tematik

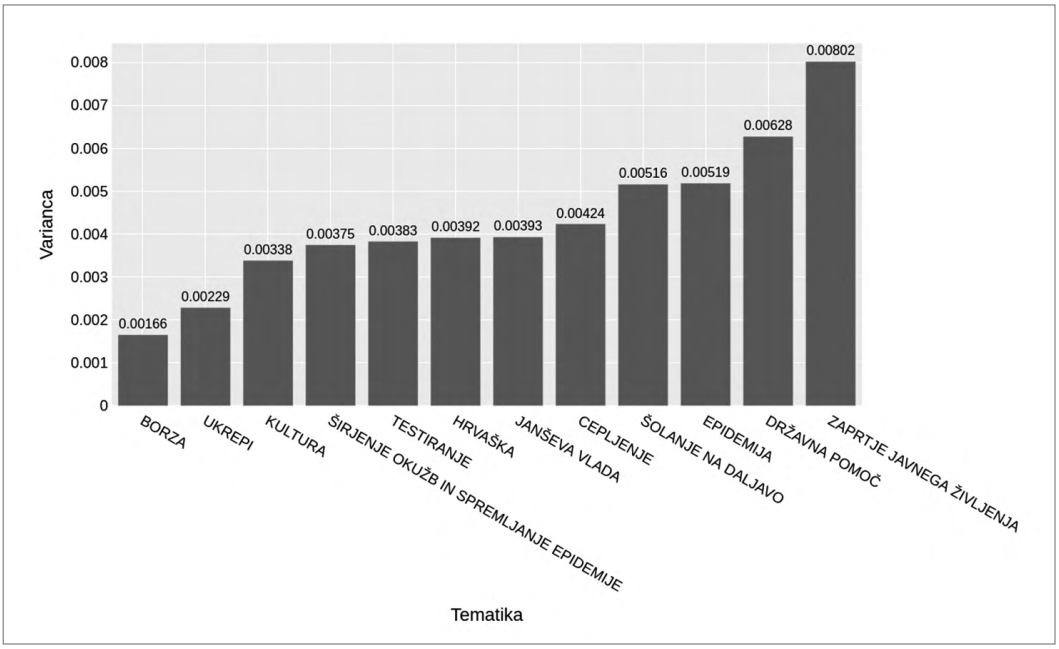
Za vektorsko reprezentacijo tematike za posamezen medij smo povprečili vse vložitve lem, ki tematiko opisujejo (besede v tematiki pripadajočem oblaku na Sliki 1). Z izračunom variance na množici reprezentacij na nivoju posameznega medija za vsako specifično tematiko lahko pridobimo oceno, kako se razlikuje poročanje o posamezni temi. Bolj kot se konteksti, v katerih se skupina besed iz tematike pojavlja, razlikujejo med mediji, večja bo varianca.

Iz Slike 2 vidimo, da se med štirinajstimi mediji najbolj raznoliko poroča o zaprtju in državni pomoči, najbolj enovito pa je poročanje o borzni tematiki. Raznolikost poročanja bi lahko bila povezana z različnimi zornimi koti in poudarki, ki jih imajo mediji na različne teme, vendar bi bilo za natančnejše razumevanje povezave med variancami in vsebinami potrebno nadaljne raziskovanje.

2. Pri metodi fastText je vsaka beseda predstavljena kot vsota vektorskih vložitev znakovnih n-gramov, ki jih beseda vsebuje. V praksi to pomeni, da metoda pri modeliranju semantične bližine upošteva tudi morfološko podobnost besed, zaradi česar je ta metoda še posebej uporabna za generiranje besednih vložitev v morfološko bogatih jezikih, kot je slovenščina.

3. Ta metoda za izdelavo kontekstualnih vložitev temelji na nevronski arhitekturi Transformer (Vaswani in dr. 2017), ki uporablja mehanizem pozornosti za določanje semantičnih relacij med besedami v kontekstu. Model, ki smo ga uporabili, je bil naučen na nenadzorovan način, na nalogi napovedovanja maskiranih žetonov v slovenskem korpusu, ki vsebuje 3,5 milijarde besed. Pri tej nalogi se 15% žetonov v korpusu zamenja z maskiranimi žetoni, model pa se nauči napovedovanja teh maskiranih žetonov s pomočjo nezamaskiranega konteksta.

Slika 2: Varianca tematik v medijih Korpusa-14.



Povezanost konceptov s covid-19

Z računanjem povezanosti med besedami (s pomočjo kosinusne razdalje med njihovimi vektorji) primerjamo povezanost izbranih konceptov s covid-19 v različnih portalih Korpusa-14.

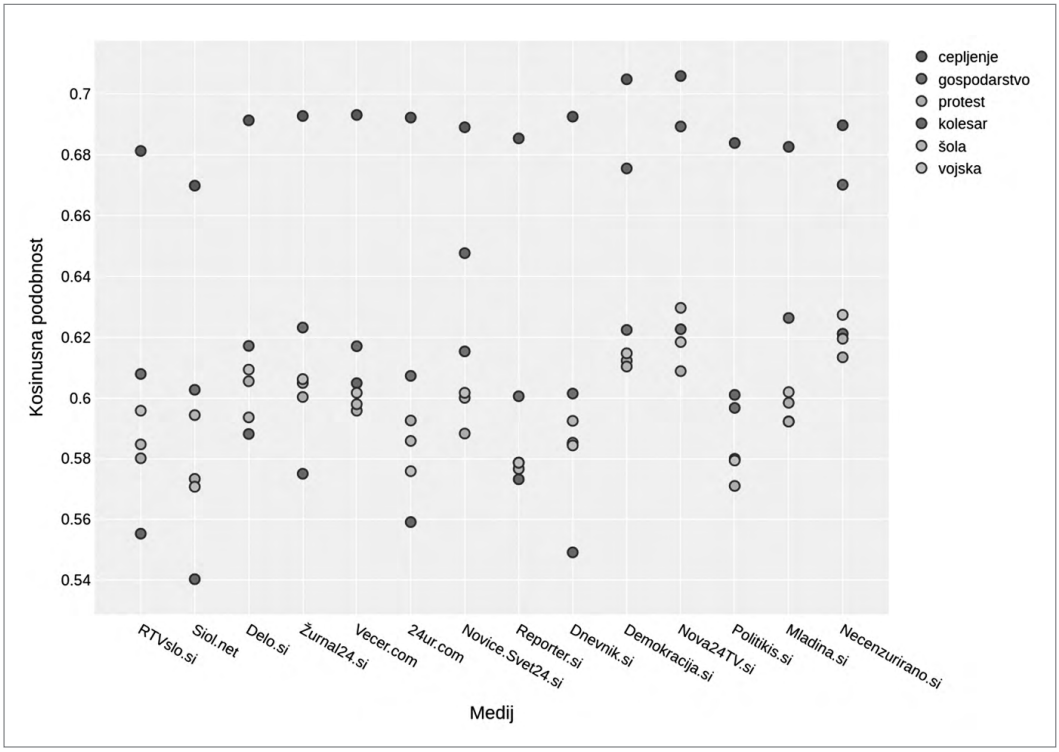
Besede so bile izbrane ročno glede na kriterij nudenja vpogleda v več aspektov epidemije in poročanja s strani različnih medijev. Kot zelo očitno povezano besedo smo izbrali *cepljenje*. Ker nas je zanimalo, ali se je več poročalo o gospodarskih ali izobraževalnih posledicah epidemije, smo vključili besede *gospodarstvo* in *šola*, dodali smo besedo *vojska*. Prav tako smo izbrali besede, za katere smo predpostavljali večje razlike med mediji (*kolesar* in *protest*).

Iz Slike 3 je razvidno, da je na vseh portalih najmočnejša povezava med covid-19 in cepljenjem, kar je pričakovano. Zanimiva je beseda *gospodarstvo*, ki se pri večini portalov bolj povezuje s covid-19 kot na primer beseda šola, kar morda priča o tem, da mediji v kontekstu epidemije covid-19 večji poudarek dajejo gospodarskim temam kot šolstvu, četudi so se zdele javne razprave o izvajanju šolanja v času epidemije prav tako v ospredju kot teme vezane na gospodarstvo.

Z vidika primerjave med portali se največja odstopanja pojavijo pri besedah, ki se manj samoumevno pojavljajo v povezavi s covid-19. To so hkrati tudi besede, za katere bi lahko rekli, da imajo večjo "ideološko obteženost", ker v javnih razpravah pogosto nastopajo kot označevalci s polariziranimi ideološkimi pomeni. To ponazarja primer besede *kolesar*, ki se močneje povezuje s covid-19 na portalih nova24tv.si, demokracija.si in necenzurirano.si. Veliko manj izrazita pa je povezava med besedo *kolesar* in covid-19 na siol.net. Povezanost besede kolesar s covid-19 je mogoče razložiti s politizacijo besede *kolesar*, tako pri medijih, ki so družbeni iniciativi kolesarjev naklonjeni (necenzurirano.si) kot tistimi, ki skušajo iniciativo diskreditirati (nova24tv.si, demokracija.si), manj pa tam, kjer je beseda rabljena izrazito v športnem kontekstu. Beseda *kolesar* je namreč v času epidemije, ko se je kolesarjenje

vzpostavilo kot protivladno protestniško gibanje, dobila nove konotativne pomene (konotira junaški upor proti vladni represiji na eni strani, na drugi pomeni razdiralno gibanje, ki škoduje aktualnemu političnemu establišmentu).

Slika 3: Povezave med izbranimi koncepti in covid-19 po različnih medijih.

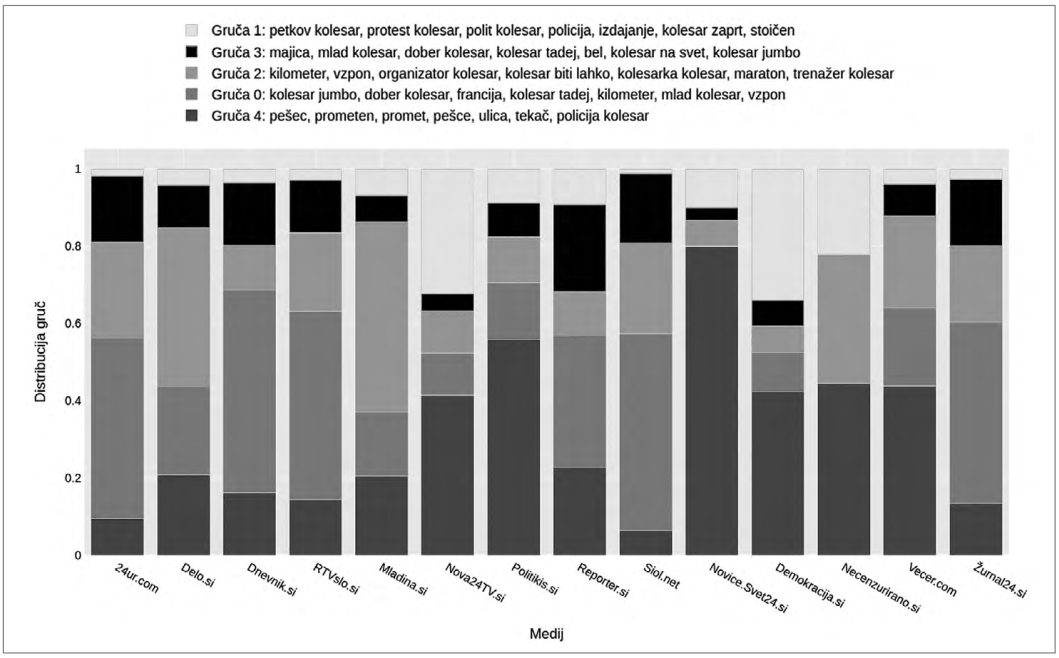


Razlikovanje besednih rab

Podobno kot Martinc in dr. (2021) smo za analizo razlik med mediji uporabili metodo gručenja pomenov in primerjavo distribucije gruč. Kontekstualne vektorje besednih pojavitev gručimo s pomočjo algoritma k-means (Steinley 2006). Za vsako besedo v Korpusu-14 njene rabe razdelimo na 5 gruč. Vsako gručo opišemo s skupkom ključnih besed oz. besednih nizov glede na TF-IDF. Razlike med portali lahko nato preučujemo z vidika razlik med distribucijami gruč.

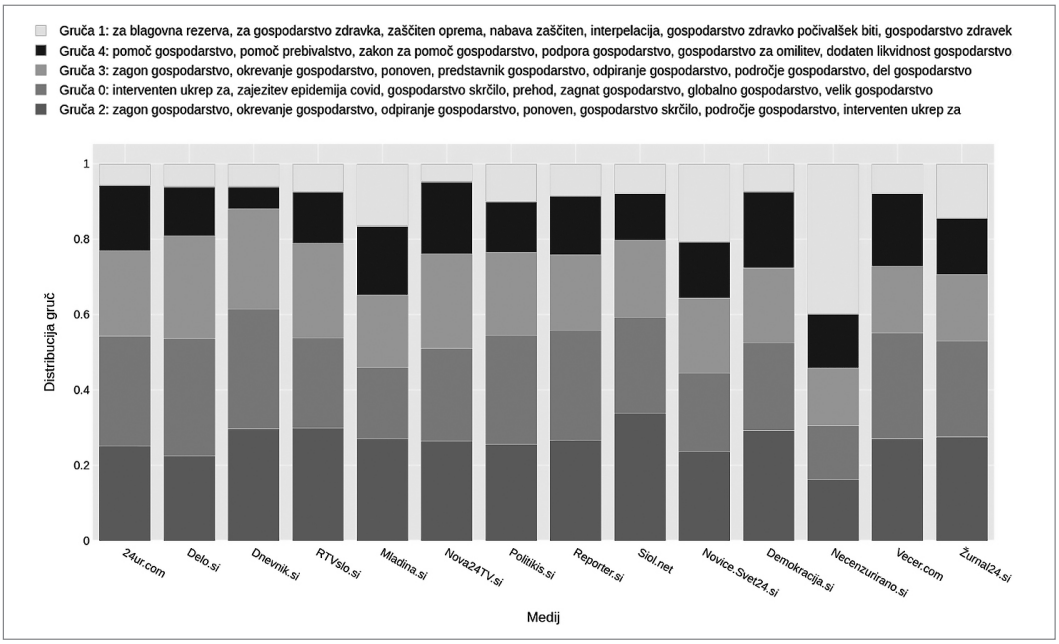
Metodo ponazorimo na primeru besede *kolesar* (Slika 4). Različne rabe besede *kolesar* v različnih gručah opisujejo pripadajoče besede. Zanimiva je predvsem gruča 1, ki se nanaša na politično konotacijo besede, saj jo označujejo pojmi kot *petkov kolesar, protest kolesarjev, policija*. Ostale štiri gruče pa se nanašajo na rabe besede kolesar v drugih, predvsem športnih kontekstih. Iz razlik med distribucijami lahko vidimo, da je gruča 1 izrazitejše zastopana na portalih demokracija.si, nova24tv.si in v nekoliko manjši meri v necenzurirano.si.

Slika 4: Prikaz gruč za besedo kolesar.



Tudi pri analizi gruč za besedo *gospodarstvo* (Slika 5) je zanimiva gruča 1 (z besedami blagovna rezerva, zaščitna oprema in interpelacija), ki zaznamuje ključne sporne teme v javnosti. Gruča 1 je najbolj zastopana v necenzurirano.si, ki je v času epidemije te teme tudi najbolj izpostavljal v kontekstu kritike delovanja vlade. Ta gruča je izrazitejša tudi na portalu mladina.si, ki se prav tako izrazito postavlja kot kritik vladnega delovanja, ter pri novičarskem tabloidu novice.svet24.si.

Slika 5: Prikaz gruč za besedo gospodarstvo.



Zaključki

V prispevku s pomočjo računalniških metod analiziramo poročanje slovenskih novičarskih portalov o epidemiji covid-19. Z metodo LDA identificiramo osrednje teme poročanja, kot so *epidemija, državna pomoč, šolanje na daljavo, cepljenje, gospodarstvo*, idr. Nato z uporabo kontekstualnih besednih vložitev analiziramo razlike v poročanju izbranih portalov. Zanimive razlike so predvsem pri bolj „ideološko obteženih" besedah, kot je *kolesar*, kjer je povezava s covid-19 močnejša tako na portalih, ki so protestom izraziteje naklonjeni (necenzurirano.si) kot med tistimi, ki skušajo iniciativo diskreditirati (nova24tv.si, demokracija.si). Prav tako z analizo različnih kontekstualnih pomenov pokažemo, da je na teh portalih politična raba besede *kolesar* bistveno bolj zastopana. V nadaljevanju bi bilo zanimivo pogledati tudi druge besede, ki polarizirajo javno razpravo (npr. *cepljenje, migrant, meja*).

Zahvala

Prispevek je rezultat raziskovalnega projekta Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami (št. J6-2581) in programa Tehnologije znanja (št. P2-0103), ki ju financira ARRS, ter evropskega projekta EMBEDDIA (No. 825153), ki ga v okviru okvirnega programa za raziskave in inovacije Obzorje 2020 financira EU. Predstavljeni izsledki ne predstavljajo mnenja Evropske komisije in predstavlja izključno mnenja avtorjev.

Literatura

Blei, David M., Ng, Andrew. Y., in Jordan, Michael. I. (2003): Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993–1022.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, in Mikolov, Tomas (2017): Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5: 135–146.

Cantril, Hadley (1999): Invazija z Marsa. V: S. Splichal (ur.): Komunikološka hrestomatija: 137–150. Ljubljana: Fakulteta za družbene vede.

Entman, Robert M. (1994): Representation and Reality in the Portrayal of Blacks on Network Television News. Journalism Quarterly, 71(3): 509–520.

Hart, Sol P., Chinn, Sedona, Soroka, Stuart (2020): Politicization and Polarization in COVID-19 News Coverage. Science communication, 42(5): 679–697.

Leban, Gregor, Fortuna, Blaž, Brank, Janez, in Grobelnik, Marko (2014): Event Registry: Learning about World Events from News. V: Proceedings of the 23rd International Conference on World Wide Web (WWW ,14 Companion): 107–110. New York: Association for Computing Machinery.

McCombs, Maxwell E., in Shaw, Donald L. (1972): The Agenda-Setting Function of Mass Media. The Public Opinion Quarterly, 36(2): 176–187. Dostopno prek: http://www.jstor.org/stable/2747787 (14. 6. 2021).

Hubner, Austin (2021): How did we get here? A framing and source analysis of early COVID-19 media coverage. Communication Research Reports, 38(2): 112–120

iPROM in Valicon (2021): Medijska potrošnja 2021. Dostopno prek: https://iprom.si/files/2021/05/iPROM-in-Valicon-raziskava-Medijska-potrosnja-2021-Porocilo-iPROM-Press.pdf (15. 6. 2021).

Liu, Qian, Zheng, Zequan, Zheng, Jiabin, Chen, Qiuyi, Liu, Guan, Chen, Sihan, Chu, Bojia, Zhu, Hongyu, Akinwunmi, Babatunde, Huang, Jian, Zhang, Casper J. P., in Ming, Wai-Kit (2020):

Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. Journal of Medical Internet Research, 22(4): e19118

Martinc, Matej, Kralj Novak, Petra, in Pollak, Senja (2020): Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. V: Proceedings of the 12th Conference on Language Resources and Evaluation: 4811–4819.

Martinc, Matej, Perger, Nina, Pelicon, Andraž, Ulčar, Matej, Vezovnik, Andreja, in Pollak, Senja (2021): EMBEDDIA Hackathon Report: Automatic Sentiment and Viewpoint Analysis of Slovenian News Corpus on the Topic of LGBTIQ+. V: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation: 121–126.

Montariol, Syrielle, Martinc, Matej, in Pivovarova Lidia (2021): Scalable and Interpretable Semantic Change Detection. V: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 4642–4652.

Mutua, Sylvia Ndanu, in Ong'ong'a, Daniel Oloo (2020): Online News Media Framing of COVID-19 Pandemic: Probing the Initial Phases of the Disease Outbreak in International Media. European Journal of Interactive Multimedia and Education, 1(2): e02006.

Rebello, Katarina, Schwieter, Christian, Schliebs, Marcel, Joynes-Burgess Kate, Elswah, Mona, Bright, Jonathan, in Howard, N. Philip. (2020): Covid-19 News and Information from State-Backed Outlets Targeting French, German and Spanish-Speaking Social Media Users. Understanding Chinese, Iranian, Russian and Turkish Outlets. Data memo. Dostopno prek: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/06/Covid-19-Misinfo-Targeting-French-German-and-Spanish-Social-Media-Users.pdf (3. 6. 2021).

Steinley, Douglas (2006): K-Means Clustering: a Half-Century Synthesis. British Journal of Mathematical and Statistical Psychology, 59(1): 1–34.

Ulčar, Matej, in Robnik-Šikonja, Marko (2020): Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0, Slovenian language resource repository CLARIN.SI. Dostopno prek: http://hdl.handle.net/11356/1387 (1. 5. 2021).

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion in Gomez, Aidan N., Kaiser, Lukasz in Polosukhin, Illia (2017): Attention is all you need. V: Proceedings of the 31st International Conference on Neural Information Processing Systems: 6000–6010.

Appendix E: Knowledge Graph informed Fake News Classification via Heterogeneous Representation Ensembles

Knowledge Graph informed Fake News Classification via Heterogeneous Representation Ensembles

Boshko Koloski
Jožef Stefan Int. Postgraduate School
Jožef Stefan Institute
1000 Ljubljana

Timen Stepišnik Perdih
Jožef Stefan Institute
1000 Ljubljana

Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
1000 Ljubljana

Senja Pollak
Jožef Stefan Institute
1000 Ljubljana

Blaž Škrlj
Jožef Stefan Int. Postgraduate School
Jožef Stefan Institute
1000 Ljubljana

Abstract

Increasing amounts of freely available data both in textual and relational form offers exploration of richer document representations, potentially improving the model performance and robustness. An emerging problem in the modern era 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. In this work we explore how different document representations, ranging from simple symbolic bag-of-words, to contextual, neural lan-

*Fully documented templates are available in the elsarticle package on CTAN.

guage 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 (grounded) subject-predicate-object triplets. We demonstrate that knowledge graph-based representations already achieve competitive performance to conventionally accepted representation learners. Furthermore, when combined with existing, contextual representations, knowledge graph-based document representations can achieve 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.

Keywords: fake news detection, knowledge graphs, text representation, representation learning, neuro-symbolic learning
2010 MSC: 00-01, 99-00

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 [1]. In the times of a global pandemic they can endanger the global health, for example by reporting that using bleach can stop the spread of Coronavirus [2, 3], or that vaccines are problematic for human health. With the upbringings of the development of the information society, the increasing capability to create and spread news in various formats makes the detection of problematic news even harder.

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.

By being able to process large collections of labeled and unlabeled textual inputs, contemporary machine learning approaches are becoming a viable solution to automatic e.g., credibility detection [4]. One of the key problems, however, 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 that are freely accessible to contextual language models capable of differentiating between subtle differences between a multitude of texts [5]. This work explores how such technologies can be used to aid and prevent spreading of problematic content, at scale.

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 [6], this path has not yet been sufficiently explored for fake news identification. The main contributions of this work, which significantly extend our conference paper [7] 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 remaining work is structured as follows. In Section 2, we present the relevant related work, followed by the text and graph representations used in our study in Section 3, we present the proposed method, followed by the evaluation

in Section 4. We discuss the obtained results in Sections 5 and 6 and finish with the concluding remarks in Sections 7 and 8.

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 [8]. Other fake news related tasks include the identification of a potential author as a spreader of fake news and the verification of facts. Many of the contemporary machine learning approaches are based on deep neural-network models [9].

Despite the fact that the neural network based approaches outperform other approaches on many tasks, they are not directly **interpretable**. On the other side, 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. For example, a recent approach explored document representation enrichment with symbolic knowledge (Wang et. al [10]). 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. [6] 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. was

preformed by Koloski et al. [11], 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.

Versatile approaches achieve state of the art results when considering various tasks related to fake news detection; Currently, the transformer architecture [12] is commonly adopted for various down-stream learning tasks. The winning solution to the COVID-19 Fake News Detection task [13] utilized fine-tuned BERT model that considered Twitter data scraped from the COVID-19 period - January 12 to April 16, 2020 [14, 9]. Other solutions exploited the recent advancements in the field of Graph Neural Networks and their applications in these classification tasks [15]. However, for some tasks best preforming models are SVM-based models that consider more traditional n-gram-based representations [16]. Interestingly, the stylometry based approaches were shown [17] to be a potential threat for the automatic detection of fake news. 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 [18]. The popularity of deep learning and the successes of Convolutional and Recurrent Neural Networks motivated development of models following these architectures for the tasks of headline and text matching of an article [19]. Lu and Li [20] 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. 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.

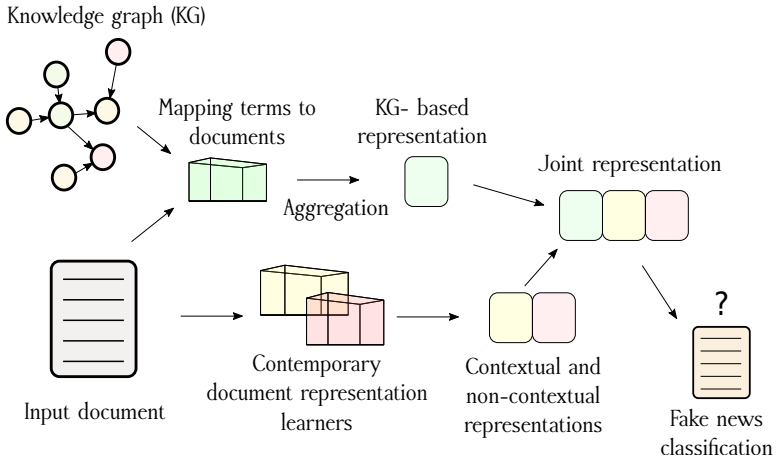


Figure 1: Schematic overview of the proposed methodology. Both knowledge graph-based features and contextual and non-contextual document features are constructed, and used simultaneously for the task of text classification.

3. Proposed methodology

In this section 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 1. 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 ex-

ploiting the methods we already developed in our submission to the COVID-19 fake news detection task [7]. We next discuss the document representations considered in this work.

Hand crafted features. We use stylometric features inspired by early work in authorship attribution [8]. 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, standard deviation of the word length in document. Additionally we counted the number of words beginning with upper and the number of words beginning 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 Koloski et al. [21] solution to the PAN 2020 shared task on Profiling Fake News Spreaders on Twitter [22] 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[23]. Next, we used the prepared data for feature construction. For the feature construction we used the technique used by Martinc et al. [24] 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 [25] 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.

Contextual features. For capturing contextual features we utilize embedding methods that rely on the transformer architecture [12], including:

- DistilBert [26] *distilbert-base-nli-mean-tokens* - d = 768 dimensions
- RoBERTa [27] - *roberta-large-nli-stsb-mean-tokens* - d = 768 dimensions
- XLM [28] - *xlm-r-large-en-ko-nli-ststb* - d = 768 dimensions

First, we applied the same preprocessing as described in subsection 3.1. 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 intermediary layers [29].

3.2. Knowledge graph-based document representations

We continue the discussion by presenting the key novelty of this work: 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 multi-graph, 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 useful manner. To obtain the representations we use six knowledge graph embedding techniques: TransE [30], RotatE[31], QuatE[32], ComplEx[33], DistMult[34] and SimplE[35]. The goal of a knowledge graph embedding method is to obtain numerical representation of the KG, or in the case of this work, 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 1. It can be observed that RotatE is the only method capable of modeling all five relations.

Table 1: Relations captured by specific knowledge graph embedding from the GraphVite knowledge graph suite [36].

Name	Symmetry	Anti-symmetry	Inversion	Transitivity	Composition
TransE [30]	x	x	✓	✓	x
DistMult [34]	✓	x	x	x	x
ComplEx [33]	✓	✓	✓	✓	x
RotatE [31]	✓	✓	✓	✓	✓
QuatE [32]	✓	✓	✓	✓	x
SimplE [35]	✓	✓	✓	✓	x

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 learn the space separately and later combine them in order to obtain the higher dimensional 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 [37] (Figure 2). The most central nodes include terms such as 'encyclopedia' and 'united state'.

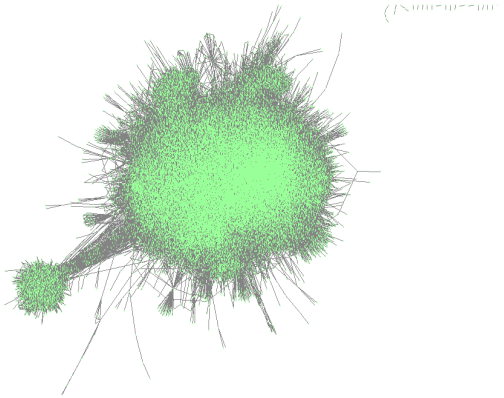


Figure 2: The WikiData5m knowledge graph - the ≈100,000 most connected nodes. It can be observed that multiple smaller structures co-exist as part of the global, well connected structure.

The GraphVite library [36] 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 WordNetLemma-*

tizer tool.

In the obtained texts, we search for concepts (token sets) consisting of unigrams, 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:

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

The considered aggregation scheme, albeit being 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 work 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 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 potential of the multi-modal representations we consider

Table 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

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 Subsection 3.2 - TransE TransE, SimplE, ComplEx, QuatE, RotatE and DistMult. We agreggate the concepts with the AGGAVERAGE strategy.

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

- LM+KG - we combine the induced KG representations with the methods explained in Subsection 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 model 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* or *hinge* loss function. We applied the learners on the three different representations scenarios.

Representation stacking with neural networks. Since we have various representations both for the textual patterns and for the embeddings of 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. We tested three different neural networks for learning this task.

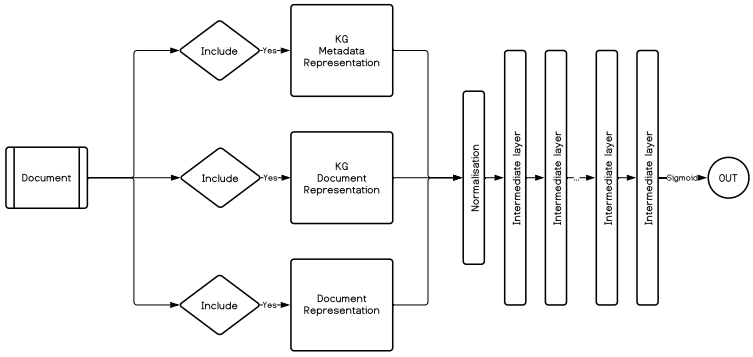


Figure 3: Neural network architecture for learning the joint intermediate representations. The *Include* decision block implies that some of the representations can be optionally excluded from the learning. The number of the intermediate layers and the dimensions are of varying sizes and are part of the model's input.

The proposed architecture consists of main two 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 [7]. 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 activation layer with the number of unique outputs.

4. Empirical evaluation

In this section, 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 multilabel fake news detection problem. We next discuss the data sets related to each problem considered in this work.

COVID-19 Fake News detection data set [13, 38] 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 three different sets: train, validation and test.

Liar, Liar Pants on Fire [39] represents a subset of PolitiFact's collection of news that are labeled in 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 with six different degrees of a fake news provided. For each news article, an additional metadata is provided consisting of: speaker, controversial statement, US party to which the subject belongs, what the text address and the occupation of the subject.

Profiling fake news Spreaders is an author profiling task that was organized under the PAN2020 workshop [22]. 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 [40] 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 3.

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%)

Table 3: Distribution of samples per given label in the three splits: train, validation and test for all four data sets respectively.

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 use if 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 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 regularization. For the SGD learner we defined a custom hyperparameter 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 [41]. We used the *SELU* 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 [7].
- 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.

5. Quantitative results

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

5.1. 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 the top performing our model. The metadata-entity based representation outperforms the induced representations by a margin of 2.42%, this is due the captured relations between the entities from the metadata. The evaluation of the data is task with respect to the models is shown in Table 4.

Table 4: 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 continually 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) [42]	**0.3740**	x	x	x

5.2. Task 2: FakeNewsNet

The Log(2) neural network 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 rate 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 methodology improves the score over the current best performing model by a margin of 3.22%. The evaluation of the data is task with respect to the models is shown in Table 5.

Table 5: 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) [43]	0.8558	x	x	x

5.3. 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 learn the intermediate representations more successfully due to the lack of data examples(only 300 were provided). The addition of factual knowledge (embedded with the TransE and RotatE methods) to the text representation improved the score of the model

improving the LM based representation by 10% gain in accuracy, and 8.59% gain in F1-score.

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(only 300 examples provided for the training). However, the problem benefited 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 evaluation of the data is task with respect to the models is shown in Table 6.

Table 6: 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-GradientDescent 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) [16]	**0.7500**	x	x	x

5.4. 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 meta-data information, so we ommited the KG-ENITTY evaluation. The evaluation of the data is task with respect to the models is shown in Table 7.

Table 7: 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) [9]	x	**0.9869**	x	x

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 best-performing models were the ones that utilized the textual representations and the factual knowledge of concepts appearing in the data.

6. Qualitative results

In the following section we further explore the constructed multi-representation space. In Subsection 6.1, 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 6.3, we analyze whether predictions can be explained with the state-of-the-art explanation methods.

6.1. 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 work we considered mutual information-based ranking [44], 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 4. 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 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. 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

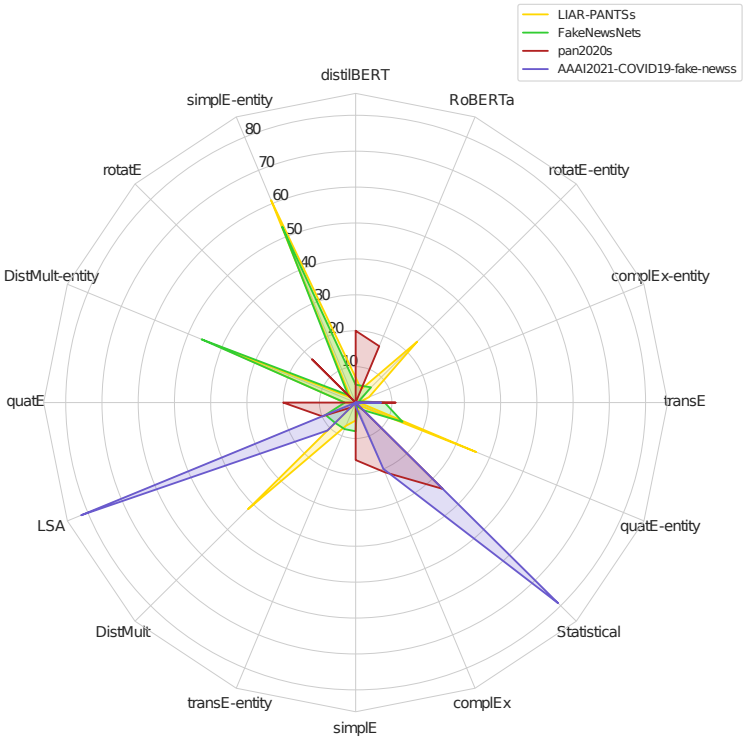
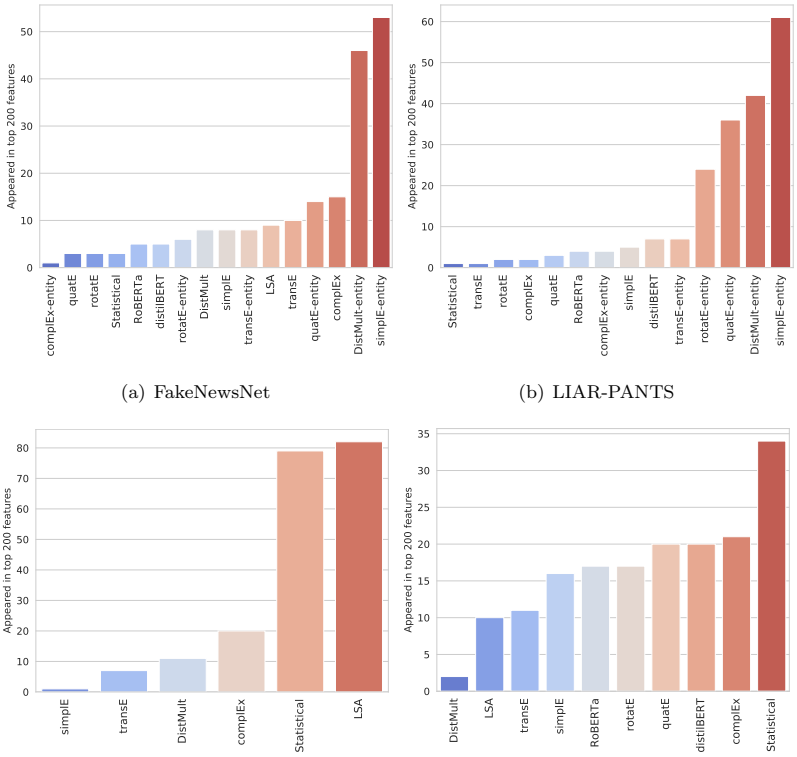


Figure 4: Overview of the most relevant feature subspaces for individual data sets.

utility of multiple representations and their different 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.



(a) FakeNewsNet (b) LIAR-PANTS

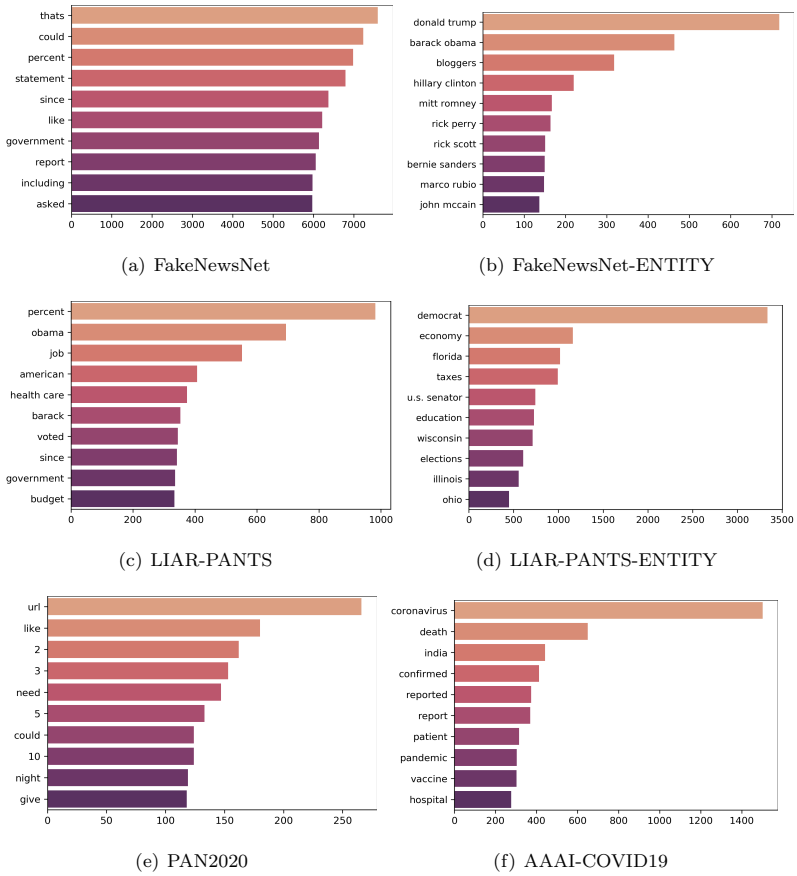
(c) AAAI-COVID19 (d) PAN2020

Figure 5: 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.

6.2. 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* dataset. The retrieved concepts are shown in Figure 6.



(a) FakeNewsNet (b) FakeNewsNet-ENTITY

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

(e) PAN2020 (f) AAAI-COVID19

Figure 6: 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.

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 describing the state of the author such as *need, give, could, and like*. Examples of tweets with present words are given in Appendix A.

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 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, 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 the Appendix in Figure B.9. 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. Additional distribution details are given in Appendix B.

6.3. 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 [45] and generalized them using Reasoning with Explanations (ReEx) [46] to discover their hypernyms, which can serve as human understandable explanations. Figure 7 shows words with the highest variance in their respective class, while Figure 8 shows found hypernyms of words with the highest variance for each of the classes.

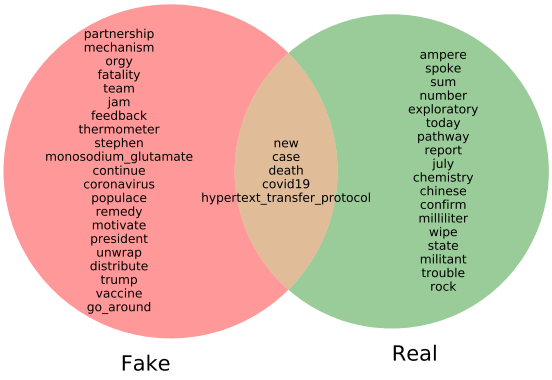


Figure 7: Words with the highest variance in their class. This is the first step towards providing understandable explanations of what affects the classification.

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".

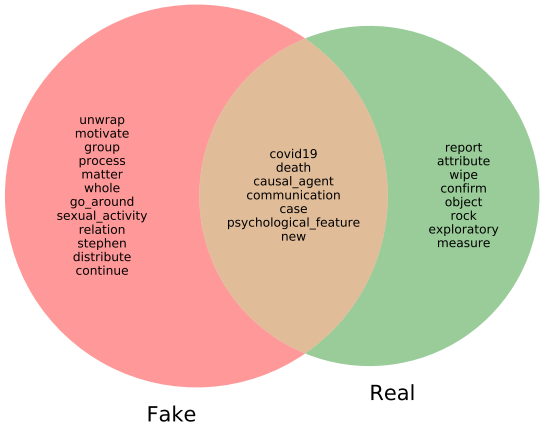


Figure 8: We used ReEx with Wordnet to generalize words with the highest variance in their class, and produce understandable explanations.

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.

7. Discussion

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 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. Wang et al. [10] 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.

Different knowledge embedding methods capture different relational properties. For this study we performed a combination with models that covered Symmetry, Anti-symmetry, Inversion, Transitivity and Composition property. 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.

8. Conclusions

We compared different representations methods for text, graphs and concepts, and proposed a novel method for merging them into a more efficient representation for 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 tackled tasks.

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 this scalability we propose exploring some dimensionality-reduction approaches such as UMAP [47] that map the original space to a low-dimensional manifold. Another problem of the method is choosing the right approach for concept extraction from a given text. 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 as further work further improve the mapping quality and with it the resulting representations.

For further work we propose exploring attention based mechanisms to derive explanations for the feature significance of a classification of an instance. Additionally we would like to explore how the other aggregation methods such as the *AGG-TF* and the *AGG-TF-IDF* perform on the given problems. The intensive amount of research focused on the Graph Neural Networks represents another potential field for exploring our method. The combination of different KG embedding approaches captures different patterns in the knowledge graphs.

The code is freely accessible at `https://gitlab.com/boshko.koloski/codename_fn_b/`.

Acknowledgments

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors also acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103 and P6-0411), the project CANDAS (No. J6-2581), the CRP project V3-2033 as well as the young researcher's grant of the last author.

References

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

[2] C. M. Pulido, L. Ruiz-Eugenio, G. Redondo-Sama, B. Villarejo-Carballido, 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). doi:10.3390/ijerph17072430.
URL https://www.mdpi.com/1660-4601/17/7/2430

[3] A. B. Kadam, S. R. Atre, Negative impact of social media panic during the COVID-19 outbreak in India, Journal of Travel Medicine 27 (3), taaa057. arXiv:https://academic.oup.com/jtm/article-pdf/27/3/taaa057/33245047/taaa057.pdf, doi:10.1093/jtm/taaa057.
URL https://doi.org/10.1093/jtm/taaa057

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

[5] K. Shu, S. Wang, D. Lee, H. Liu, Mining disinformation and fake news: Concepts, methods, and recent advancements, in: Disinformation, Misinformation, and Fake News in Social Media, Springer, 2020, pp. 1–19.

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

[7] B. Koloski, T. S. Perdih, S. Pollak, B. Škrlj, Identification of covid-19 related fake news via neural stacking, arXiv preprint arXiv:2101.03988.

[8] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 231–240. doi:10.18653/v1/P18-1022.
URL https://www.aclweb.org/anthology/P18-1022

[9] A. Glazkova, M. Glazkov, T. Trifonov, g2tmn at constraint@ aaai2021: Exploiting ct-bert and ensembling learning for covid-19 fake news detection, arXiv preprint arXiv:2012.11967.

[10] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph and text jointly embedding, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1591–1601. doi:10.3115/v1/D14-1167.
URL https://www.aclweb.org/anthology/D14-1167

[11] B. Koloski, B. Škrlj, M. Robnik-Šikonja, Knowledge graph-based document embedding enrichment.
URL https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=119701

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, 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, 2017, pp. 5998–6008.
URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[13] P. Patwa, S. Sharma, S. PYKL, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, arXiv preprint arXiv:2011.03327.

[14] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: Anatural language processing model to analyse covid-19 content on twitter.

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

[16] J. Buda, F. Bolonyai, 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.
URL http://ceur-ws.org/Vol-2696/

[17] T. Schuster, R. Schuster, D. J. Shah, R. Barzilay, The limitations of stylometry for detecting machine-generated fake news, Computational Linguistics 46 (2) (2020) 499–510. doi:10.1162/coli_a_00380.
URL https://www.aclweb.org/anthology/2020.cl-2.8

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

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

[20] Y.-J. Lu, C.-T. Li, GCAN: Graph-aware co-attention networks for explainable fake news detection on social media, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 505–514. doi:10.18653/v1/2020.acl-main.48.
URL https://www.aclweb.org/anthology/2020.acl-main.48

[21] B. Koloski, S. Pollak, B. Škrlj, Multilingual detection of fake news spreaders via sparse matrix factorization, in: CLEF, 2020.

[22] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, 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, 2020.
URL CEUR-WS.org

[23] E. Loper, S. Bird, NLTK: The natural language toolkit, in: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 63–70. doi:10.3115/1118108.1118117.
URL https://www.aclweb.org/anthology/W02-0109

[24] M. Martinc, B. Skrlj, S. Pollak, 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, Vol. 2125 of CEUR Workshop Proceedings, CEUR-WS.org, 2018.
URL http://ceur-ws.org/Vol-2125/paper_156.pdf

[25] N. Halko, P.-G. Martinsson, J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions (2009). arXiv:0909.4061.

[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108. arXiv:1910.01108.
URL http://arxiv.org/abs/1910.01108

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692. arXiv:1907.11692.
URL http://arxiv.org/abs/1907.11692

[28] A. Conneau, G. Lample, 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, 2019, pp. 7057–7067.
URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html

[29] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
URL https://www.aclweb.org/anthology/D19-1410

[30] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, 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. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 2787–2795.
URL https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

[31] Z. Sun, Z. Deng, J. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.
URL https://openreview.net/forum?id=HkgEQnRqYQ

[32] S. Zhang, Y. Tay, L. Yao, Q. Liu, 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, 2019, pp. 2731–2741.
URL https://proceedings.neurips.cc/paper/2019/hash/d961e9f236177d65d21100592edb0769-Abstract.html

[33] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, 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, Vol. 48 of JMLR Workshop and Conference Proceedings, JMLR.org, 2016, pp. 2071–2080.
URL http://proceedings.mlr.press/v48/trouillon16.html

[34] B. Yang, W. Yih, X. He, J. Gao, L. Deng, 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,

2015.
URL http://arxiv.org/abs/1412.6575

[35] S. M. Kazemi, D. Poole, 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, 2018, pp. 4289–4300.
URL https://proceedings.neurips.cc/paper/2018/hash/b2ab001909a8a6f04b51920306046ce5-Abstract.html

[36] Z. Zhu, S. Xu, J. Tang, M. Qu, 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, ACM, 2019, pp. 2494–2504. doi:10.1145/3308558.3313508.
URL https://doi.org/10.1145/3308558.3313508

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

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

[39] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426.

doi:10.18653/v1/P17-2067.
URL https://www.aclweb.org/anthology/P17-2067

[40] F. S. A. J. B. H. Amirkhani, Fnid: Fake news inference dataset (2020). doi:10.21227/fbzd-sw81.
URL https://dx.doi.org/10.21227/fbzd-sw81

[41] D. P. Kingma, J. Ba, 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, 2015.
URL http://arxiv.org/abs/1412.6980

[42] T. Alhindi, S. Petridis, S. Muresan, Where is your evidence: Improving fact-checking by justification modeling, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 85–90. doi:10.18653/v1/W18-5513.
URL https://www.aclweb.org/anthology/W18-5513

[43] A. Bidgoly, H. Amirkhani, F. Sadeghi, Fake news detection on social media using a natural language inference approach.

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

[45] Princeton University, About wordnet.

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

[47] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426.

Appendix A. Examples of real and fake tweets

In this section we present some examples of real and fake tweets with words present (bold).

Appendix A.1. Real

- **fatality**,#IndiaFightsCorona: India's Total Recoveries continue to rise cross 32.5 lakh today 5 Statffes contribute 60% of total cases 62% of active cases and 70% of total fatality reported in India #StaySafe #IndiaWillWin https://t.co/KRn3GOaBNp
- **team**,An important part of our work is data collection and analysis At 11:30pm every day our data Team collates results received from all testing laboratories to inform Nigerians of the number of new #COVID19 cases Results not received at this time are reported the next day https://t.co/Nyo6NeImRk
- **partnership**,Finally we launched the first real version of the COVID Racial Data Tracker in partnership with @DrIbram and the @AntiracismCtr. This has been a major effort by our project's volunteers—and we hope it will be useful to communities across the country. https://t.co/hTyV0MA5tA team,In @followlasg our rapid response team is working with NFELTP to strengthen community testing for #COVID19 in LGAs. The team provides support to newly reactivated LGA walk-in testing sites for increased testing capacity access and awareness of #COVID19 at the grassroot level. https://t.co/MnIu3OBT3v
- **fatality**,#IndiaFightsCorona Health Ministry reviews COVID Management & Response in 15 districts across 5 States exhibiting high caseload and fatality.

Appendix A.2. Fake

- **state**,India has lost over 50000 individuals to coronavirus till date. In view of the rising coronavirus cases Bihar government extends lockdown

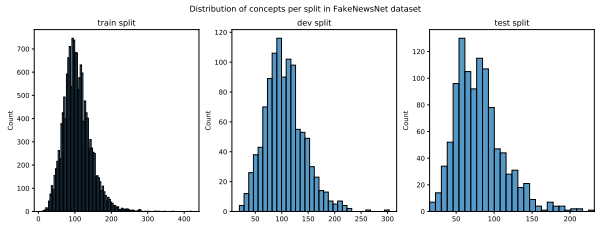
in the state till 6 September. At Nationalist Congress Party chief Sharad Pawar's residence four people tested positive for #coronavirus. https://t.co/LqGJHHVr2g

- **report**,Leaked Report Says There Are T́oo Many HumansÓn The Planet https://t.co/03kvl3oOXU #globalwarming #coronavirus #conspiracy
- **today**,"???Covid is never going away! This is the beach today in Raleigh, North Carolina.??"
- **report**,"In an Aaj Tak news report the Chinese prime minister said ""Reading Quran and offering namaz is the only cure for COVID-19."""
- **chinese**,"In an Aaj Tak news report the Chinese prime minister said ""Reading Quran and offering namaz is the only cure for COVID-19."""

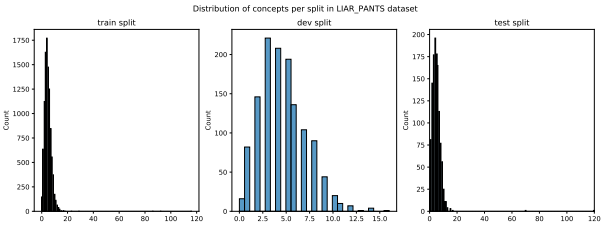
Appendix B. Distribution of concepts

In this subsection we showcase the distribution of concepts per each data set, shown in Figure B.9.

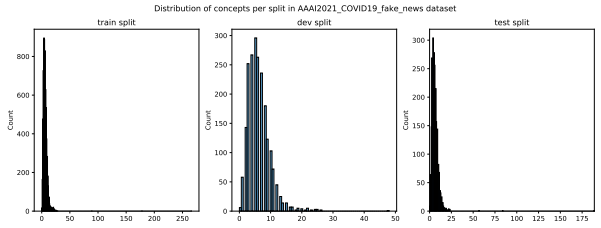
Figure B.9: Distribution of concepts extracted from the WikiData5m KG per article in the data sets.



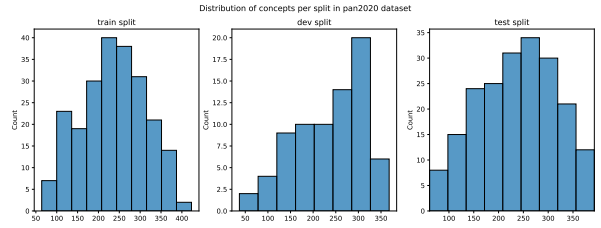
(a) FakeNewsNet



(b) LIAR-PANTS



(c) AAAI-COVID19



(d) PAN2020

Appendix C. Performance of individual feature spaces

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

Appendix C.1. Evaluation of all subsets of spaces

In this subsection 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 dataset where we use the whole dataset, due to the small number of examples. For every problem we evaluate all the possible combinations consisted of KG representatiosn and LM represetntations, 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.

Appendix C.1.1. 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 C.10. The worst-performing combinations are listed in Table C.8, while the best-performing combinations are listed in Table C.9.

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 C.8: Liar worst 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

Table C.9: LIAR best 10 representation combinations.

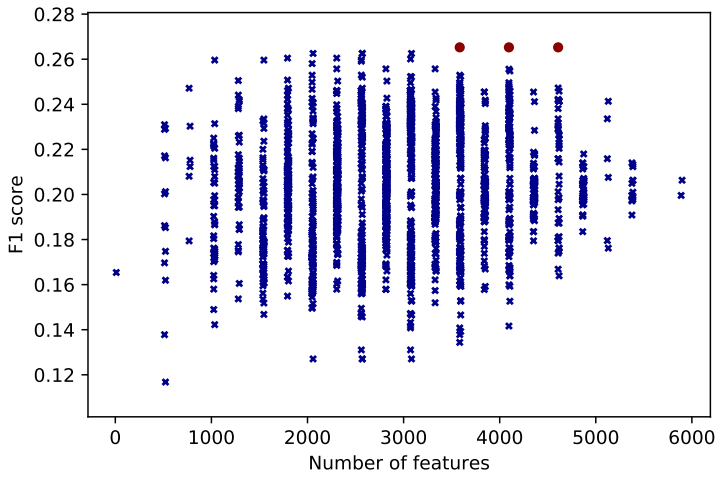


Figure C.10: The interaction of dimensions and the F1-score for the LIAR problem. The red dots represent the highest scoring models.

Appendix C.1.2. FakeNewsNet

Knowledge graph and their combinations generated too general spaces that scored lowest on the dataset. The lowest scoring representation is the one based only on the TransE KG embedding method. Notable improvement was seen with introduction of the contextual 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 C.11. The worst-performing combinations are listed in Table C.10, while the best-performing combinations are listed in Table C.11.

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 C.10: FakeNewsNet worst 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

Table C.11: FakeNewsNet best 10 representation combinations.

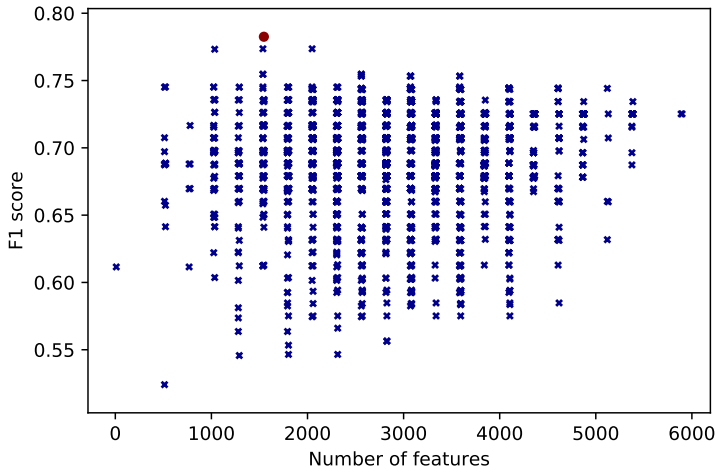


Figure C.11: The interaction of dimensions and the F1-score for the FakeNewsNet problem. The red dots represent the highest scoring models.

Appendix C.1.3. 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 statsitical language features, with a dimension of 1546. The worst-performing combinations are listed in Table C.12, while the best-performing combinations are listed in Table C.13. The dependence of the number of features and the f1-scores is represented in Figure C.12.

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 C.12: PAN2020 worst 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

Table C.13: PAN2020 best 10 representation combinations.

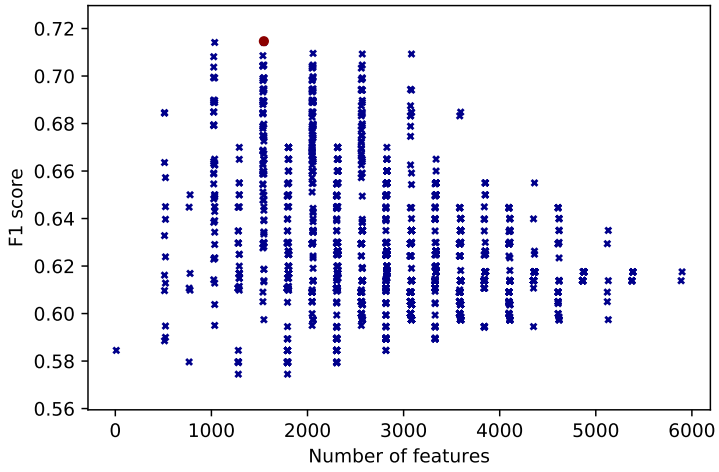


Figure C.12: The interaction of dimensions and the F1-score for the PAN2020 problem. The red dots represent the highest scoring models.

Appendix C.1.4. COVID-19

Knowledge graph only based representation yielded too general spaces, making for the lowest-performing spaces for the COVID-19 task. Notable improvement for the dataset was achieved by the addition of language models to the knowledge graph representations. The worst-performing combinations are listed in Table C.14, while the best-performing combinations are listed in Table C.15. The dependence of the number of features and the f1-scores is represented in Figure C.13.

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 C.14: COVID-19 worst 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

Table C.15: COVID-19 best 10 representation combinations.

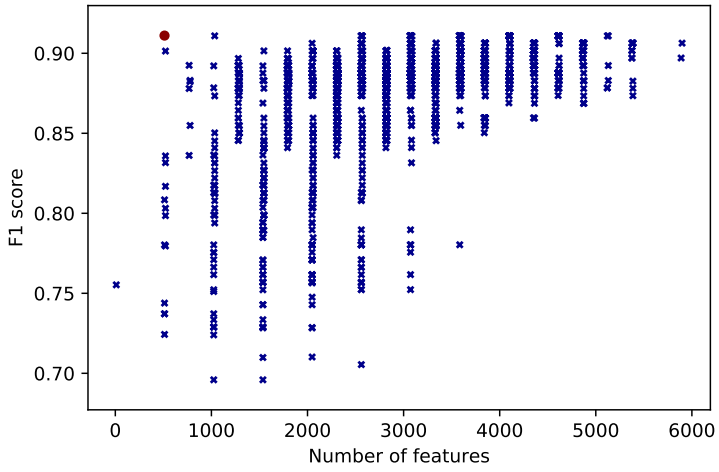


Figure C.13: The interaction of dimensions and the F1-score for the COVID-19 problem. The red dots represent the highest scoring models.

Appendix C.2. Conclusion

In this section we discuss the main highlights of the extensive ablation studies targeting the performance of different feature space combinations. The main conclusions are as follows.

In the evaluation of spaces study, we analyzed how combining various spaces before learning common joint spaces impacts performance. We can take two different outputs from the study:

1. knowledge graph-based representations on their own are too general for tasks where the main type of input are short texts. However, including 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.

Appendix F: Grammatical Profiling for Semantic Change Detection

Grammatical Profiling for Semantic Change Detection

Mario Giulianelli* ILLC, University of Amsterdam m.giulianelli@uva.nl
Andrey Kutuzov* University of Oslo andreku@ifi.uio.no
Lidia Pivovarova* University of Helsinki first.last@helsinki.fi

Abstract

Semantics, morphology and syntax are strongly interdependent. However, the majority of computational methods for semantic change detection use distributional word representations which encode mostly semantics. We investigate an alternative method, grammatical profiling, based entirely on changes in the morphosyntactic behaviour of words. We demonstrate that it can be used for semantic change detection and even outperforms some distributional semantic methods. We present an in-depth qualitative and quantitative analysis of the predictions made by our grammatical profiling system, showing that they are plausible and interpretable.

1 Introduction

Lexical semantic change detection has recently become a well-represented field in NLP, with several shared tasks conducted for English, German, Latin and Swedish (Schlechtweg et al., 2020), Italian (Basile et al., 2020) and Russian (Kutuzov and Pivovarova, 2021a). The overwhelming majority of solutions employ either static word embeddings like word2vec (Mikolov et al., 2013) or more recent contextualised language models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These models build upon the distributional semantics hypothesis and can capture lexical meaning, at least to some extent (e.g., Iacobacci et al., 2016; Pilehvar and Camacho-Collados, 2019; Yenicelik et al., 2020). Thus, they are naturally equipped to model semantic change.

Yet it has long been known for linguists that semantics, morphology and syntax are strongly interrelated (Langacker, 1987; Hock and Joseph, 2019). Semantic change is consequently often accompanied by morphosyntactic drifts. Consider the English noun '*lass*': in the 20th century, its 'SWEETHEART' meaning became more dominant over the older sense of 'YOUNG WOMAN'. This was accompanied by a sharp decrease in plural usages ('*lasses*'), as shown in Figure 1.

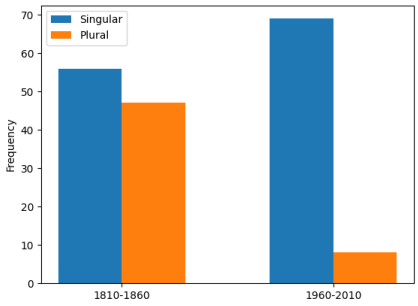


Figure 1: Changes in the number category distribution for the English noun '*lass*' over time, calculated on the English corpora of the SemEval 2020 shared task 1 (Schlechtweg et al., 2020). '*Lass*' is annotated as semantically changed in the SemEval dataset.

Exploiting distributions of *grammatical profiles*—i.e., morphological and syntactic features—to detect lexical semantic change is the focus of this paper. We investigate to what extent lexical semantic change can be detected using *only* morphosyntax. Our main hypothesis is that significant changes in the distribution of morphosyntactic categories can reveal useful information on the degree of the word's semantic change, even without help from any lexical or explicitly semantic features.

Due to the interdependence of semantics and morphosyntax, it is often difficult to determine which type of change occurred first, and whether it triggered the other. Establishing the correct causal direction is outside the scope of this study; it is sufficient for us to know that semantic and morphosyntactic changes often co-occur.

By proposing this functionalist approach to lexical semantic change detection, we are not aiming at establishing a new state-of-the-art. This

*Equal contribution, the authors listed alphabetically.

is hardly possible without taking semantics into account. But what exactly *is* possible in such a functionalist setup?

We investigate this question experimentally[1] using standard semantic change datasets for English, German, Swedish, Latin, Italian and Russian. Our main findings are the following:

1. Tracing the changes in the distribution of dependency labels, number, case, tense and other morphosyntactic categories outperforms count-based distributional models. In many cases, prediction-based distributional models (static word embeddings) are outperformed as well. This holds across six languages and three different datasets.
2. Morphological and syntactic categories are complementary: combining them improves semantic change detection performance.
3. The categories most correlated with semantic change are language-dependent, with number being a good predictor cross-linguistically.
4. The predictions derived from grammatical profiling are usually interpretable (as in the '*lass*' example above), which is not always the case for methods from prior work based on word embeddings, either static or contextualised. This makes our method suitable for linguistic studies that require qualitative explanations.

2 Related work

Behavioural profiles were introduced in corpus linguistics by Hanks (1996) as the set of syntactic and lexical preferences of a word, revealed by studying a large concordance extracted from a corpus. The behavioural profile of a word consists of corpus counts of various linguistic properties, including morphological features, preferred types of clauses and phrases, collocates and their semantic types (Gries and Otani, 2010). Subtle distinctions in word meaning are reflected in behavioural profiles. Indeed this technique, which combines lexical and grammatical criteria for word sense distinction, was used to study synonymy and polysemy (Divjak and Gries, 2006; Gries and Divjak, 2009) as well as antonymy (Gries and Otani, 2010).

One of the theoretical roots for profiling is the theory of *lexical priming* (Hoey, 2005). According to this theory, words trigger a set of grammatical and lexical constraints, referred to as *primings* and stored in a mental concordance. The theory states that 'Drifts in priming ... provide a mechanism for temporary or permanent language change' (Hoey, 2005, p. 9), and since primings are thought to be organised in the mental concordance in the form of behavioural profiles (Gries and Otani, 2010), it is theoretically plausible that diachronic word meaning change is reflected in a change of behavioural profiles. As far as we are aware, this idea has not been further developed in corpus linguistics.

In spite of its theoretical validity, behavioural profiling as a practical data analysis technique has serious limitations. Profiles include a large variety of word properties and some of them, especially those related to semantics, cannot be easily extracted from a corpus automatically. Usually, a particular subset of word properties is selected based on researchers' intuition and background knowledge, and statistical tests are sometimes used for feature selection at later stages of the analysis (Divjak and Gries, 2006). Moreover, the variety of properties comprised in a behavioural profile makes statistical analysis difficult due to correlations between language phenomena of different levels and sparsity of the data (Kuznetsova, 2015, section 2.2.2). For these reasons, some studies (Janda and Lyashevskaya, 2011; Eckhoff and Janda, 2014) reduce a word's possibly very broad behavioural profile to a more compact *grammatical profile*, i.e. a set of preferred morphological forms for the word. These studies too, however, rely on an *a priori* selection of relevant morphological tags.

These technical difficulties may explain why profiling has not been used in computational approaches to lexical semantic change detection. Most attempts to tackle word meaning change in NLP are based on distributional patterns of *lexical* co-occurrences, starting from early count-based approaches (Juola, 2003; Hilpert and Gries, 2008), continuing with dimensionality reduction techniques (Gulordava and Baroni, 2011), and later accelerated by embeddings-based models (Kutuzov et al., 2018). More recently, contextualised embeddings were also applied to this task (Giulianelli et al., 2020; Montariol et al., 2021).

As far as we are aware, there is one exception to this trend: Ryzhova et al. (2021) employed grammatical profiles to detect the semantic change of Russian nouns. In their work, a profile of case and number frequency distributions is collected separately for each time period, and the degree of

[1] Our code is available at https://github.com/glnmario/semchange-profiling

semantic change is measured as the cosine distance between the two distributions. The results obtained with this method are close to the results yielded by word2vec embeddings, but lower than those of contextualised embeddings. Inspired by Ryzhova et al. (2021), we further investigate the ability of grammatical profiles to capture word meaning change. We propose a number of improvements and evaluate them on datasets in six different languages. Most importantly, we use *all* available morphological tags, without any manual pre-selection, and we conduct an in-depth analysis of our results to understand why grammatical profiling works for this task and what are its limitations.

3 Data and tasks

Following the standard evaluation approach adopted for automatic lexical semantic change detection, we cast the problem as either binary classification (Subtask 1, using the terminology of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020)) or as a ranking task (Subtask 2). In Subtask 1, given a set of target words, a system must determine whether the words lost or gained any senses between two time periods. In Subtask 2, a system has to rank a set of target words according to the degree of their semantic change.

Annotating data for word meaning change detection is a non-trivial process because it requires taking into account numerous word occurrences from every time period of interest. The current practice adopted in the community is to annotate pairs of sentences containing a target word used either in the same or in a different sense; then pairwise scores are aggregated to obtain a final measure of change, either binary or continuous (Schlechtweg et al., 2018). This procedure has been used by organizers of three recent shared tasks: the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020), EvaLita (Basile et al., 2020) and RuShiftEval (Kutuzov and Pivovarova, 2021a). We use the data from these three shared tasks, allowing to compare our approach with the state-of-the-art results obtained by distributional models.

The SemEval dataset consists of target words in four languages—37 English, 48 German, 40 Latin, and 32 Swedish—that are manually annotated for both subtasks. The EvaLita dataset consists of 18 Italian words annotated for Subtask 1 only. Finally, the RuShiftEval dataset consists of 99 Russian nouns annotated for Subtask 2. All datasets are accompanied by diachronic corpora. Most of the corpora are split in two time periods, except for the RuShiftEval corpus, which is separated into three time bins: *Russian1* and *Russian2* are annotated with semantic shifts between the pre-Soviet and Soviet periods, and between the Soviet and post-Soviet periods respectively; *Russian3* is annotated with semantic shifts between the pre-Soviet and post-Soviet periods (Kutuzov and Pivovarova, 2021b).

In sum, we have at our disposal several dozens words from three Indo-European language groups: Italic, Germanic and Slavic. Though our results may not generalize to other language families or to other languages within the families analysed, these are the most diverse data that are currently available for this kind of study.

4 Methods

4.1 Basic procedure

To obtain grammatical profiles, the target historical corpora are first tagged and parsed with UDPipe (Straka and Straková, 2017).[2] Then we count the frequency of morphological and syntactic categories for each target word in both corpora. More precisely, we count the FEATS values of a corpus's CONLLU file and store the frequencies in two data structures—one for each time period. For example, `{'Number=Sing': 338, 'Number=Plur': 114}` is the morphological dictionary obtained for an English noun in a single time period. We store syntactic features in an additional dictionary, where keys correspond to the labels of the dependency arc from the target word to its syntactic head (as found in the DEPREL field of a CONLLU-formatted corpus).

For each target word and for both morphological and syntactic dictionaries, we create a list of features by taking the union of keys in the corresponding dictionaries for the two time bins. The feature list will be `['Number=Sing', 'Number=Plur']` for the example above. Then, we create feature vectors $\vec{x}_1$ and $\vec{x}_2$, where each dimension represents a grammatical category and the value it takes is the frequency of that category in the corresponding time period. If a feature does

[2] We use the following models: *english-lines-ud-2.5*, *german-gsd-ud-2.5*, *latin-proiel-ud-2.5*, *swedish-lines-ud-2.5*, *russian-syntagrus-ud-2.5*, *italian-isdt-ud-2.5*.

not occur in a time period, its value is set to 0. The resulting feature vectors represent grammatical profiles for a word in the corresponding periods. Since the feature list is produced separately for each word, the size of the vectors varies across words.

Finally, we compute the cosine distance $cos(\vec{x}_1, \vec{x}_2)$ between the vectors to quantify the change in the grammatical profiles of the target word. This is done separately for morphological and syntactic categories, yielding two distance scores d_{morph} and d_{synt}. They are used directly to rank words in Subtask 2: the larger is the distance, the stronger is the semantic change. To solve the binary classification task (Subtask 1), we classify the top n target words in the ranking as 'changed' (1) and the rest of the list as 'stable' (0). The value of n can be either set manually or inferred from the ranking using off-the-shelf algorithms of change point detection (Truong et al., 2020).

We also combine the scores obtained separately for morphological and syntactic tags by averaging d_{morph} and d_{synt} for each target word (rounding to the nearest integer in the case of binary classification) and then re-rank the words according to the resulting values. In the end, we have three solutions for each task: 'morphology', 'syntax' and 'averaged'. In the next subsections, we describe a number of improvements that we use to amend this basic procedure.

4.2 Filtering

To reduce noise that could be introduced due to rare word forms and possible tagging errors, we exclude rare grammatical categories from the analysis. A feature is filtered out from a feature vector $\vec{x}$ if the sum of the feature occurrences in the two time slices amounts to less than five percent of the total word usages. It is possible to optimise this threshold, but we do not tune any numerical parameters to avoid over-fitting to the target datasets.

4.3 Category separation

In the basic procedure described above, we extract exactly one morphological feature for each word occurrence; this type of morphological feature is a combination of morphological categories that exhaustively describes a word form. For example, this is an excerpt from a grammatical profile of the English verb '*circle*' in the 1810-1860 time period:

```
Tense=Pres|VerbForm=Part : 50
Mood=Ind|Tense=Past|VerbForm=Fin : 24
Tense=Past|VerbForm=Part|Voice=Pass : 17
VerbForm=Inf : 9
Mood=Ind|Tense=Pres|VerbForm=Fin : 1
Tense=Past|VerbForm=Part : 1
```

This representation is very sparse—some features appear only once in the corpus—and it conflates categories of different nature, such as verb form and tense. We therefore introduce a category separation step, where feature vectors are created separately for each morphological category. Thus, we transform a distribution of *word forms* into a distribution of *morphological categories* and obtain a denser and more meaningful representation:

```
Tense : {Past 42, Pres 51}
VerbForm : {Part 68, Fin 25, Inf 9}
Mood : {Ind 25}
Voice : {Pass : 17}
```

Then cosine distance is computed for each category separately. In the example above, we obtain separate distance values for `Tense`, `VerbForm`, `Mood`, and `Voice`; the number of distances differs across words and languages. We take the maximum distance value as the final change score, assuming that a significant change in the distribution of a single category indicates semantic change, regardless of the other categories.[3]

When separation is combined with filtering, filtering is performed *after* feature separation to preserve maximum information. Continuing with the previous example: in the basic procedure, the word form `Tense=Past|VerbForm=Part` is filtered out, as it appears once in the first corpus and it is rare in the second corpus as well. In the category separation strategy this form is taken into account, separately contributing to the `Tense` and `VerbForm` distances.

4.4 Combination of morphology and syntax

Category separation opens new possibilities for taking syntactic categories into account. We can average morphological and syntactic distances, as in our basic procedure, or append the syntactic distance value to the array of morphological distances, and then choose the maximum. In the first strategy, morphological and syntactic rankings are weighted equally regardless of the number of morphological categories for a given word. In the second strategy,

[3] We also experimented with averaging category distances. This improves the results compared to using categories without separation, but it is not as effective as taking the maximum.

syntactic labels are weighted down depending on the richness of the morphological profile.

5 Results

We evaluate our method on both subtasks of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task (Schlechtweg et al., 2020). As described in Section 3, Subtask 1 is a binary classification task, evaluated with accuracy. Subtask 2 is a ranking task, evaluated with Spearman's rank correlation.

Basic procedure Using only morphological features, we obtain an average correlation of 0.181 across the four SemEval languages, as can be seen in Table 1. Syntactic features yield a +0.017 increase, and after averaging d_{morph} and d_{synt} (see Section 4.1) we reach a correlation score of 0.208. This is already substantially higher than the SemEval baseline which employed count-based distributional models (see Table 1).

Frequency threshold Filtering out rare features as described in Section 4.2 has a small but positive impact on all three setups: +0.011 for morphological features, +0.033 for syntactic features, and +0.065 for the combination of the two.

Category separation Measuring distance between morphological categories separately (see Section 4.3) produces an additional significant boost: we obtain a correlation score of 0.278 using these refined morphological representations. In combination with syntactic features (Section 4.4), this approach yields an average correlation of 0.369 with human judgements. This is our best result on Subtask 2, more than twice higher than a correlation obtained by the SemEval count-based baseline (see Table 1); for Latin, a language with rich morphology, grammatical profiles actually outperform even the *best* SemEval 2020 submission. These scores are particularly impressive given that, unlike those based on distributional vectors, our method has no access to lexical semantic information.

As can be seen in Table 1, our category separation approach does not extend well to the Russian test sets, obtaining an average correlation score of 0.130.[4] A possible explanation for the lower correlation may be related to smaller distances between Russian time bins as compared to the SemEval setup: *Russian1* and *Russian2* are annotated with semantic shifts between pre-Soviet and Soviet and between Soviet and post-Soviet periods respectively, while *Russian3* measures the change between pre-Soviet and post-Soviet periods, with a significant time gap in between. Indeed we obtain much higher scores on *Russian3*. In addition, the annotation procedures for the RuShiftEval dataset differ in some details from those for SemEval'20.

Another observation is that morphological category separation does not improve results for English. The best method for English relies only on syntactic features. The most plausible explanation is that English morphology is rather poor and it tends to mark grammatical categories with separate words. Our method can be potentially improved by taking into account multi-word forms, e.g. to determine English verb mood.

Subtask 1 Following our basic procedure (Section 4.1), we assign a classification score of 1 to the top 43% of the target words[5] for each language, ranked according to their grammatical profile changes. This yields an accuracy close to that of the SemEval count-based baseline (see Table 2).[6] Filtering rare features hardly yields any improvement here, but once combined with morphological category separation and automatic change point detection it produces an accuracy of 0.603. We also observe that using change point detection with dynamic programming (Truong et al., 2020) does not cause any significant accuracy decrease in comparison to using the hard-coded 43% ratio, showing that our method does not require knowledge of the test data distribution. On the Italian test set, we correctly classify 3 more words (out of 18) than the collocation-based baseline (Basile et al., 2019b), obtaining an accuracy of 0.778.

6 Qualitative analysis

In Section 5, we showed that grammatical profiling alone can detect a word meaning change better than count-based distributional semantic models which exploit lexical co-occurrence statistics. This is a remarkable finding: it confirms that meaning change leaves traces in grammatical profiles and it demonstrates that these traces can be used as effective predictors of a word's meaning stability. In this Section, to better understand when change in grammatical profiles is a good indicator of lexical

[4] At the same time, in the basic procedure, morphological features yield a much higher correlation score of 0.225.

[5] Average ratio of changed words across SemEval datasets.

[6] Note that the SemEval'20 count baseline also uses a manually defined threshold value in Subtask 1.

Categories	SemEval 2020 languages					Russian			
	English	German	Latin	Swedish	Mean	Russian1	Russian2	Russian3	Mean
	Basic procedure								
Morphology	0.234	0.043	0.241	0.207	0.181	**0.137**	0.210	**0.327**	**0.225**
Syntax	0.319	0.163	0.328	-0.017	0.198	0.060	0.101	0.269	0.143
Average	0.293	0.147	0.304	0.088	0.208	0.101	0.191	0.294	0.195
	5% filtering								
Morphology	0.211	0.080	0.285	0.191	0.192	0.127	0.185	0.264	0.192
Syntax	**0.331**	0.146	0.265	0.184	0.231	0.056	0.111	0.279	0.149
Average	0.315	0.171	0.345	0.263	0.273	0.094	0.183	0.278	0.185
	Category separation and 5% filtering								
Morphology	0.218	0.074	0.519	0.303	0.278	0.028	**0.241**	0.293	0.187
Average	0.321	0.227	0.523	**0.381**	0.363	0.002	0.179	0.278	0.153
Combination	0.320	**0.298**	**0.525**	0.334	**0.369**	0.000	0.149	0.242	0.130
Prior SemEval results						**Prior RuShiftEval results***			
Count baseline	0.022	0.216	0.359	-0.022	0.144	0.314	0.302	0.381	0.332
Best shared task system	0.422	0.725	0.412	0.547	0.527	0.798	0.803	0.822	0.807
(Ryzhova et al., 2021)	-	-	-	-	-	0.157	0.199	0.343	0.233

Table 1: Performance in graded change detection (SemEval'20 Subtask 2 and RuShiftEval), Spearman rank correlation coefficients. Note that RuShiftEval features three test sets for three different time period pairs.
*The RuShiftEval baseline relies on CBOW word embeddings and their local neighborhood similarity. (Ryzhova et al., 2021) used an ensemble method with much higher performance, we report the results obtained solely with profiling. While SemEval results are fully unsupervised, the best RuShiftEval results are supervised and not directly comparable to our setting.

Categories	English	German	Latin	Swedish	Mean	Italian
	Basic procedure					
Morphology	0.595	0.521	0.525	0.581	0.555	0.722
Syntax	0.541	**0.646**	0.575	0.645	0.602	0.611
Average	0.568	0.583	0.475	**0.710**	0.584	0.722
	Automatic change point detection					
Morphology	**0.622**	0.479	**0.625**	0.548	0.569	0.722
Syntax	0.514	0.625	0.500	0.677	0.579	0.611
Average	0.595	0.542	0.525	0.677	0.585	**0.778**
	Category separation, change point detection and 5% filtering					
Morphology	0.622	0.583	**0.625**	0.581	**0.603**	0.500
Average	0.595	0.625	0.450	**0.710**	0.595	0.667
Combination	0.541	0.583	0.575	0.645	0.586	0.500
Prior SemEval results						**Prior EvaLita results***
Baseline	0.595	0.688	0.525	0.645	0.613	0.611
Best shared task system	0.622	0.750	0.700	0.677	0.687	0.944

Table 2: Performance in binary change detection (SemEval'20 Subtask 1 and EvaLita), accuracy. Note that in this paper we mostly focus on ranking (Subtask 2). All the binary change detection methods here are entirely based on the scores produces by the ranking methods.
*The Italian baseline relies on collocations (Basile et al., 2019a): for each target word, two vector representations are built, with the Bag-of-Collocations related to the two different time periods. Then, the cosine similarity between them is computed.

semantic change, we analyse the characteristics of the target words to which our method assigns the most and least accurate rankings.

6.1 When is grammatical profiling enough?

We begin by analysing the most accurately ranked words (see Appendix A). The Italian word '*lucciola*', for example, is ranked 1st out of 18 by our method due to the singular usages of the word disappearing after 1990. The singular usage is indeed much more likely for the dying sense of the word (an euphemism for 'PROSTITUTE'), whereas the plural form '*lucciole*' is more likely used for the stable sense of the word ('FIREFLIES') or in the idiomatic expression *prendere lucciole per lanterne* (*getting the wrong end of the stick*), which makes up for most of the occurrences between 1990 and 2014. Another example of correctly identified semantically shifted words is the Latin '*imperator*' (ranked 1st out of 40). In the second time period—ranging from 0 to 2000 A.D.—nominative usages become predominant. A possible explanation for this change is that the more frequent agentive usages of the word correspond to the new role of the 'EMPEROR' in the imperial Rome (27 B.C. to A.D. 476) rather than that of a generic 'COMMANDER'—the older sense of the word.[7]

For English, the noun '*stab*' is ranked 4th out of 37, mostly because of syntactic changes: 27% of its occurrences in the 20th century are used as oblique arguments, compared to only 13% in the 19th century. This is arguably associated with the emergent sense of 'SUDDEN SHARP FEELING' ('*...left me with a sharp stab of sadness*'). The German word '*artikulieren*' correctly receives a high rank (9th out of 48): it occurs only 3 times in the 19th century and 210 times between 1946 and 1990, shifting towards a much richer grammatical profile. Sharp changes in frequency are reflected in the diversity of grammatical profiles and can also help detect lexical semantic change.

Our qualitative analysis reveals that the successful examples are often cases of broadening and narrowing of word meaning. These kinds of semantic change seem to be easily picked with profiling. However, some examples of broadening and narrowing fail to be detected, as will be shown in Section 6.2, especially if they involve metaphorical extensions of word meaning. A consistent characterisation of the kinds of semantic change detected and overlooked by our method would require diachronic corpora where both the degree and the type of semantic changes are annotated.

6.2 When it is not enough?

Although it largely outperforms simple distributional semantic models, our grammatical profiling approach is still not on par with state-of-the-art semantics-based algorithms. To find out when changes in morphosyntactic profiles are not sufficient to detect a word's meaning change, we analyse *false positives* and *false negatives*: i.e., target words that are assigned an erroneously high or low semantic change score, respectively.

False positives are words whose change in grammatical profile does not correspond to semantic change. An example of a false positive is the Italian word '*cappuccio*' ('HOOD'). The increase from 9% to 41% of plural usages causes our method to assign this word a relatively high change score—6th out of 18 (6 words are annotated as changing in the Italian dataset). Inspecting the Italian corpora, we notice that between 1945 and 1970 the word is mainly used to describe the pointed hood of the robes typically worn by Ku Klux Klan members; after 1990, the word's context of usage becomes much less narrow. The meaning of the word, however, does not change. This type of errors is, at least to a certain extent, an artifact of the source data: grammatical profiles are less accurate when the set of domains covered by a corpus is limited.

Another type of false positives is also partially related to corpus imbalance. We have seen in the previous section that sharp frequency increases correspond to significant changes in grammatical profiles, and that this information can be exploited by our method to detect changing words. However, frequency change can be an unfaithful indicator of meaning change. This is the case, for example, for the German words '*Lyzeum*' ('LYCEUM'; ranked 1st out of 48), and '*Truppenteil*' (a 'UNIT OF TROOPS'; ranked 11th), and for the Latin word '*jus*' (a 'RIGHT', the 'LAW'; ranked 4th out of 40).

False negatives, on the other hand, are words whose semantic change is not reflected in changes in grammatical profile. The German word '*ausspannen*' ('TO REMOVE', 'TO UNCLAMP') is used across the 19th and 20th century only in its infini-

[7] We are aware that the current separation of the Latin corpus into two time periods can be controversial. Still, we follow the splits defined by the SemEval 2020 organisers (Schlechtweg et al., 2020) for consistency and comparability with prior work.

tive form, so our method assigns it a relatively low change score (23rd out of 48). Most of the occurrences in the 19th century, however, are literal usages of the word (e.g., *die Pferde ausspannen*, *to unhitch the horses*), whereas in the (second part of the) 20th century the novel metaphorical usage of the word (e.g., *für fünf Minuten ausspannen*, *to relax for five minutes*) is the most frequent one. Another example of a German word whose novel metaphorical sense remains undetected (ranked 31st) is '*Ohrwurm*' ('EARWORM'): the grammatical profile of this word remains stable (except for the accusative case becoming slighlty more frequent), but the word acquires the meaning of *catchy song*, or *haunting melody*. Similarly, the singular usages of the Latin word '*pontifex*' increase from 63% to 83%, signalling the semantic narrowing of the word occurred in medieval Latin (from a 'BISHOP' to the 'POPE'), but the case distribution remains similar; this results in a rather low change score (ranked 22nd out of 40). The last two examples show that taking the maximum distance across categories (see 4.3) is a correct strategy, yet sometimes the changes in that grammatical category are still insufficient for our method to detect change.

7 Category importance

In this Section, we conduct an additional experiment to find out which grammatical categories are most related to semantic change. To this end, we train logistic regression classifiers for binary classification using English, German, Latin, Swedish and Italian data. The classifier features are cosine distances between frequency vectors of each particular category from different time bins. Before fitting the classifier, each feature is independently zero-centered and scaled to the unit variance. Then, regression coefficients are estimated for each feature: we consider positive weights as an indication of usefulness of a feature for classification. The outcome of this analysis is shown in Table 3. We list English nouns and verbs separately since the SemEval'20 dataset explicitly annotates part-of-speech tags for the English target words. This is not the case for the other languages in this dataset.

In line with the results presented in Section 5, Swedish and Italian classifiers yield the highest accuracy and F-score. Latin, a highly inflectional language, has by far the largest set of categories contributing positively to semantic change detection (interestingly, excluding syntax). English, a highly analytical language, is on the other end of the spectrum.

Language	Top categories	Accur.	F1
English nouns	number	0.576	0.523
English verbs	verb form, syntax	0.750	0.733
German	number, syntax, gender	0.542	0.541
Swedish	syntax, mood, voice, definiteness, number	0.839	0.797
Latin	voice, number, degree, case, gender, mood, aspect, person, tense	0.650	0.649
Italian	number, tense, syntax	0.778	0.723

Table 3: Categories with positive weights in binary classifiers of semantic change (logistic regression). 'Syntax' stands for dependency relation to the syntactic head of the word. Evaluation scores are calculated on the train data, F1 is macro-averaged.

Additionally, we estimate the relative importance of morphosyntactic categories by calculating the Spearman's rank-correlation of their respective cosine distance values (across all target words) with the gold semantic change rankings. In other words, we single out each category, e.g. verbal mood, and test whether diachronic change in its frequency distribution is correlated with manually annotated semantic change scores.

In Table 4, we show the categories with statistically significant ($p < 0.05$) correlations for each language and dataset. In English, as expected given its analytical nature, only changes in syntactic roles yield such a correlation; other categories are either non-existent in this language, or are not linked to semantic change strongly enough. For an inflection language such as Latin, number and adjectival degree are highly predictive (the latter is arguably because Latin has the highest ratio of adjectives among all SemEval 2020 Task 1 datasets: about 20%). Not surprisingly for a synthetic language, the morphological categories of number and case show strong correlations for Russian. In the case of the larger time gap between pre-Soviet and post-Soviet periods (Russian 3), syntactic relationships also become a good predictor.

What *is* surprising, however, is that changes in gender are also correlated with semantic change in the Russian case. This result is hard to inter-

	Number	Mood	Degree	Gender	Case	Syntax
English	-	-	-	-	-	0.331
German	-	-	-	-	-	-
Latin	0.304	-	0.301	-	-	-
Swedish	0.402	0.397	-	-	-	-
Russian 1	-	-	-	0.218	0.196	-
Russian 2	-	-	-	0.231	0.324	-
Russian 3	0.246	-	-	0.218	0.327	0.279

Table 4: Spearman rank correlations between diachronic grammatical profile distances for different categories and manually annotated semantic change estimations. '-' stands for no significant correlation.

pret, since grammatical gender is a lexical feature of Russian nouns and does not change from occurrence to occurrence; even diachronically, such cases are quite rare. The reason for this is slightly erroneous morphological tagging: our tagger mixes up homographic inflected forms, which abound in Russian, and assigns feminine gender to masculine nouns, and vice versa. The reliance on the tagger performance can be seen as a limitation of our grammatical profiling approach. However, the existence of the correlation hints that these errors are not entirely random, and their frequency is influenced by word usage: gender is ambiguous only in certain case and number combinations, and the frequency of these combinations seems to change diachronically. For example, for the form '*cheki*' ('cheques/grenade pin'), the masculine lemma licenses the accusative plural reading, while the feminine lemma licenses the genitive singular reading. Thus, even the tagger errors are in fact informative.

Interestingly, for German, no single category changes are significantly correlated with semantic change. This is in line with our weak—although still higher than the count-based baseline—results for German described above, but is somewhat surprising, given the fusional nature of the language, with its rich spectrum of inflected word forms.[8] Some peculiarities of the employed tagger model might be responsible for this finding, which should be further tested and explained in future work.

8 Conclusion

Semantic change is inextricably tied to changes in the distribution of morphosyntactic properties of words, i.e. their grammatical profiles. In this paper, we showed that tracking these changes is enough to build a semantic change detection system which, without access to any lexical semantic information, consistently outperforms count-based distributional semantic approaches to the task. Grammatical profiling yields surprisingly good evaluation scores across different languages and datasets, without any language-specific tuning. For Latin, a language with rich morphology, our methods even establish a new SOTA in Subtask 2 of SemEval'20 Task 1.

These results indicate that grammatical profiling cannot compete with state-of-the-art methods based on large pre-trained language models, since they have the potential to encode both semantics and grammar. Yet reaching the highest possible scores on the task was not our goal. Instead, the aim of our study was to demonstrate that more attention should be paid to the relation between morphosyntax and semantic change. Whether morphosyntactic and semantic features are complementary and can be successfully combined is a interesting question to be addressed in future work.

We performed an extensive quantitative and qualitative analysis of our semantic change detection methods, showing that profiling yields interpretable results across several languages. Nevertheless, we still lack an understanding of some aspects of the interaction between semantics and morphosyntax. Finding the reasons behind the relatively poor performance on some datasets, e.g. German, is an important direction for future studies.

Another interesting question is how to incorporate full dependency trees into grammatical profiles, rather than only dependency relations to the syntactic head of a word. This is particularly important for analytical languages, where grammatical markers are presented in more than one word, such as with English verb mood and aspect. Moreover, dependency structure can be crucial for languages from families other than the Indo-European, e.g. to take into account detached counters in Japanese or plural markers in Yoruba.

In light of our experimental results, we argue that grammatical profiling should become one of the standard baselines for semantic change detection.

Acknowledgements

We thank the anonymous CoNLL-2021 reviewers for their helpful comments. This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grants 770299 (NewsEye), 825153 (EMBEDDIA), and 819455 (DREAM).

[8] We computed correlations for German nouns and verbs separately, but did not find any significant correlation either.

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020) CEUR Workshop Proceedings (CEUR-WS.org)*.

Pierpaolo Basile, Annalina Caputo, Seamus Lawless, and Giovanni Semeraro. 2019a. Diachronic analysis of entities by exploiting Wikipedia page revisions. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 84–91, Varna, Bulgaria. INCOMA Ltd.

Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019b. Kronos-it: a Dataset for the Italian Semantic Change Detection Task. In *CLiC-it*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dagmar Divjak and Stefan Th Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2:23J60.

Hanne M Eckhoff and Laura A Janda. 2014. Grammatical profiles and aspect in old church slavonic. *Transactions of the Philological Society*, 112(2):231–258.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Stefan Th Gries and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics*, 57:75.

Stefan Th Gries and Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, 34:121–150.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Patrick Hanks. 1996. Contextual dependency and lexical sets. *International journal of corpus linguistics*, 1(1):75–98.

Martin Hilpert and Stefan Th Gries. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Hans Henrich Hock and Brian D Joseph. 2019. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Walter de Gruyter GmbH & Co KG.

Michael Hoey. 2005. *Lexical Priming: A New Theory of Words and Language*. Routledge.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Laura A Janda and Olga Lyashevskaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive linguistics*, 22(4):719–763.

Patrick Juola. 2003. The time course of language change. *Computers and the Humanities*, 37(1):77–96.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval: a shared task on semantic shift detection for Russian. *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Andrey Kutuzov and Lidia Pivovarova. 2021b. Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Julia Kuznetsova. 2015. *Linguistic profiles: Going from form to meaning via statistics*. Walter de Gruyter GmbH & Co KG.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Anastasiia Ryzhova, Daria Ryzhova, and Ilya Sochenkov. 2021. Detection of semantic changes in Russian nouns with distributional models and grammatical features. *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

Appendix

A Model predictions

Table 5 shows the top 10 ranked words for each of the target languages according to the semantic change score of our best model. Because three time bins are available for Russian, we show the change score estimated for the interval between first and second (1-2), second and third (2-3), as well as first and third (1-3) periods.

English	German	Latin	Swedish	Italian	Russian 1-2	Russian 2-3	Russian 1-3
gas	Lyzeum	imperator	bröllop	lucciola	blagodarnost	polosa	ambitsia
chairman	vorweisen	beatus	studie	palmare	vek	vek	nalozhenie
rag	Schmiere	regnum	motiv	tac	sobrat	zhest'	ponedelnik
stab	zersetzen	jus	krita	unico	vyzov	favorit	vyzov
ball	verbauen	adsumo	konduktör	pacchetto	brat	sobrat	blin
lass	Eintagsfliege	potestas	annandag	cappuccio	jubiley	ambitsia	polosa
prop	beimischen	licet	aktiv	egemonizzare	ambitsia	nalozhenie	khren
tip	Engpaß	sensus	granskare	brama	khren	jubiley	uglevodorod
record	artikulieren	nobilitas	bolagsstämma	campanello	uglevodorod	lishenie	lishenie
plane	voranstellen	sacramentum	färg	piovra	ponedelnik	blin	chastitsa

Table 5: The top 10 rankings obtained with our best method for all the target languages. The topmost word is the one with the highest assigned change score. Russian words are transliterated from Cyrillics to Latin script.

Appendix G: Benchmarks for Unsupervised Discourse Change Detection

Benchmarks for Unsupervised Discourse Change Detection

Quan Duong, Lidia Pivovarova and Elaine Zosa

University of Helsinki, Finland

Abstract

The main motivation for this work lies in the need to track discourse dynamics in historical corpora. However, in many real use cases ground truth is not available and annotating discourses on a corpus-level is hardly possible. We propose a novel procedure to generate synthetic datasets for this task, a novel evaluation framework and a set of benchmarking models. Finally, we run large-scale experiments using these synthetic datasets and demonstrate that a model trained on such a dataset can obtain meaningful results when applied to a real dataset, without any adjustments of the model.

Keywords

Discourse dynamic, News monitoring, Synthetic data, Quantitative evaluation, Neural network, Unsupervised, Pattern detection, Pivots detection, Sequence to Sequence

1. Introduction

Various computational methods, from keyword extraction to topic modelling, have been established to facilitate discourse analysis. However, studying *discourse dynamics*—the change in prevalence of certain topics, opinions, and attitudes over time—is a novel and challenging research area yet to be developed.

The term ”discourse” has many definitions across humanities and social disciplines; it could be understood either as a property of a corpus as a whole or a property of a single text and its structure. In this paper we treat discourse as a *corpus property*. A fine-grained structure of particular documents is irrelevant for our research question and ignored in the experiments. Discourse change can only be found in a *diachronic corpus*, i.e. corpus that contains data from several consecutive time periods.

Thus input for our methods is a collection of texts, split into multiple time periods. The task breaks up into three following **sub-tasks**:

1. to detect, whether a certain discourse in this collection is *non-stable*, e.g. increases or decreases;
2. to find *a subset of documents* that belong to this discourse;
3. to find *pivot point* in the timeseries, i.e. time points where non-stable behaviour of the discourse starts and ends.

HistoInformatics 2021 – 6th International Workshop on Computational History, September 30, 2021, online
quan.duong@helsinki.fi (Q. Duong); lidia.pivovarova@helsinki.fi (L. Pivovarova); elaine.zosa@helsinki.fi (E. Zosa)
0000-0001-6566-368X (Q. Duong); 0000-0002-0026-9902 (L. Pivovarova); 0000-0003-2482-0663 (E. Zosa)

Historical research questions are generally complex and involve a lot of uncertainty, thus the ground truth needed for quantitative evaluation is usually unavailable. Quite often research deals with a specific use case, focusing on a single non-annotated dataset without a proper split into training and test subsets. Thus, finding training and evaluation data for this task is currently not possible. As far as we know, there does not yet exist a diachronic corpora annotated with discourses.

To overcome this difficulty, we propose an evaluation framework using multiple synthetic datasets. The idea is to exploit manually assigned article categories, available in many news corpora. Distinct periods and spikes in the data could be mimicked by sampling from a certain label according to a certain pattern, while all other categories are sampled randomly. Synthetic datasets allow for training and evaluation models able to find a subset of documents that are related to the same theme and follow the pattern, without looking at the manually assigned labels. The source code for this study is available on Github, which is freely accessible for further development.[1]

2. Background

Discourse dynamics has been a topic of several multidisciplinary studies that apply NLP to historical or social science research questions. Quite often these studies lean on topic modelling [1, 2, 3, 4], though others use techniques, such as language models and clustering [5, 6]. Each of these studies deal with a complex research question, such as ”immigration discourse” or ”nation building”, and the suitability of the applied methods is assessed only qualitatively, using close reading or background knowledge of the field.

There were several attempts within the NLP field to model discourse change, by the means of unsupervised topic models, such as dynamic topic models [7, 8, 9]. However, these models are often evaluated qualitatively and as a result, the applicability of the models remains unclear especially for research questions that go beyond localizing well-known historical events in time. Any model has certain limitations, that are rarely articulated [10]; and quite often a basic LDA model is preferred to more sophisticated models [11].

Another task relevant to diachronic change is lexical semantic change detection [12, 13, 14]. In this task, manual data annotation is extremely challenging [15] and synthetic datasets are commonly used [16, 17, 18, 19].

This paper is positioned in between the aforementioned fields. The research question, automatic discourse change detection, is motivated by the needs of humanities scholars but the point of view is methodological: we propose an evaluation framework rather than investigate any particular use case. The evaluation procedure is based on extensive experiments on multiple synthetic datasets, an approach adopted from the closely related task of lexical semantic shift detection. We are unaware of any work approaching discourse dynamics from this angle and run experiments similar to ours, either in NLP or digital humanities literature.

[1] https://github.com/ruathudo/detangling-discourses

3. Synthetic Datasets

3.1. Yle News Corpus

The synthetic datasets are created from a corpus of news articles published from 2011 to 2018 by the Finnish broadcasting company Yle. The corpus is distributed through Finnish Language Bank (Kielipankki)[2] and is freely available for research use[3].

Each article belongs to one major category and one or more sub-categories. To create the synthetic dataset, we take articles that belong to well-separated major categories. We found 12 categories in the corpus that are suitable for this purpose: *autot* (cars), *musiikki* (music), *luonto* (nature), *vaalit* (elections), *taudit* (diseases), *työllisyys* (employment), *jääkiekko* (hockey), *kulttuuri* (culture), *rikokset* (crimes), *koulut* (schools), *tulipalot* (fires) and *ruoat* (food). These categories have a relatively balanced number of articles and cover distinct subjects, which is appropriate for creating a clean dataset for evaluation. However, a single article may cover several themes–this introduces additional noise in the synthetic datasets and thus a desirable property. After limiting our data to these 12 categories, we end up with a reduced corpus of 207,881 articles.

3.2. Discourse Change Patterns

The datasets for our experiments are sampled to simulate pre-defined patterns of discourse change. Each dataset consists of 100 artificial time points. For each time point, we randomly sample documents from several categories in such a way that one category follows a non-stable pattern—for example, increases over time—while all others remain stable, i.e. randomly oscillating.

We define six possible patterns of discourse behaviour across time, which are illustrated in Figure 2:

- **Up**: The number of articles belonging to a discourse starts increasing at certain time point, and grows until some later point, when it becomes stable.
- **Down**: The number of articles decreases between two time points, then becomes stable.
- **Up - Down**: The number of articles increases, then decreases, then becomes stable.
- **Down - Up**: The number of articles decreases, then increases, then becomes stable.
- **Spike Up**: The trend behaves similar to the Up-Down pattern but spikes are more steep and could appear several times
- **Spike Down**: The trend behaves similar to the previous one but in reversed way.

In addition we use a **Stable** pattern, with no significant change in discourse prevalence over time.

We randomly select one target category and then for this category randomly select one of the six non-stable patterns. For the target category, in each time point t, we sample a number of articles n so that the timeline follows a randomly selected pattern. We use 100 time points. While generating these sequences, we also randomly assign the pivot points when the non-stable

[2] http://urn.fi/urn:nbn:fi:lb-2017070501

[3] According to the license we cannot redistribute datasets derived from these data. Upon acceptance we will publish our code, which ensures reproducibility of our experiments, including dataset generation.

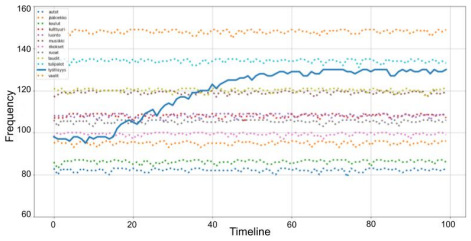


Figure 1: A sample experiment with 1 increasing category (Up) and 11 stable categories.

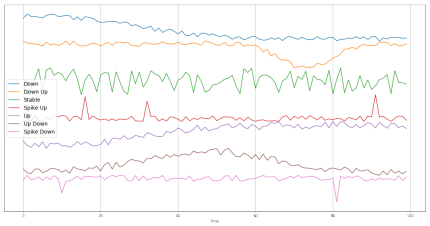


Figure 2: Seven patterns used to emulate discourse dynamics in the synthetic datasets.

pattern starts and ends, which is necessary for sub-task 3. Before and after start and end point the timeseries follows a stable pattern. Then we sample data from the remaining 11 categories, which all follow the stable pattern.

Two functions are used as basic components for discourse change: *sigmoid* or *Gaussian*. The sigmoid function is used to sample the **Up** and **Down** patterns: we assume that a novel discourse slightly increases or decreases at the beginning, then speeds up in the middle and then gradually slows up before becoming stable again, which is exactly how the sigmoid function behaves. Thus, the discourse change forms a S-curve, which is a natural shape in many language-change processes [20].

More concretely, a number of articles in Up and Down patterns follows the formula:

$$S_i = N + \frac{1}{1 + e^{-k \times (T_i - (T_{end} - T_{start})/2)}} \times N \times R$$

where T_{start} and T_{end} are the time points where the pattern starts and ends, respectively; S_i is the number of articles at time point $T_i \in [T_{start}, T_{end}]$; N is the number of articles before the starting point, R is the change rate for the pattern, arbitrarily selected between 0.3 and 0.8, and k is the parameter that defines how the change is distributed along the time. With a large k the S-curve is steep, with a slow change at two ends of the range, and a rapid change in the middle. We set $k = 0.1$ to form a gradual change.

The Gaussian function is used for the **Up - Down** and **Down - Up** patterns which have a bell shape. By modifying the mean and standard deviation of the Gaussian, we produce different forms of the bell shape, depending on the amount of data and the number of time points. We sample the bell pattern using the following formulas:

$$S = \Phi_{\mu, \sigma^2}(X) \quad \mu = (T_{end} - T_{start})/2 \quad \sigma = (T_{end} - T_{start})/k$$

$$S_i = N + \frac{T_i - \min(S)}{\max(S) - \min(S)} \times N \times R$$

The S is a set of values drawn from Probability Density Function Φ of Gaussian distribution to form a bell curve. The Gaussian distribution has μ is the middle point in time range, and σ depends on a parameter k in the equation. A large k will create a shape with a sharp peak in the middle . From our experiments, we found that $k = 5$ gives a smooth changing pattern. After

having S sampled in the bell shape, we can calculate the number of articles for each time point, however, S needs to be rescaled to have a consistent input in range $[0, 1]$ using min-max scaling as in the last equation.

Another pattern that uses the Gaussian distribution is *multiple periods* up or down spikes. This pattern will have a very short range of beginning and ending time points which is similar to a pine shape.

Figure 1 shows an example dataset: the Up pattern is used. As can be seen in the figure, random noise is added to all patterns, so small spikes are visible for all categories, including stable ones. The input to our trend-detection model are raw texts, while categories are hidden. In this way we try to emulate a realistic situation where many themes are oscillating in the news at the same time and only a few of them display a certain increasing or decreasing trend.

4. Method

In all our experiments we use two major steps: (i) building a timeseries from textual data; (ii) analysing the timeseries to classify them as either stable or unstable and finding pivot points.

We split a document collection into clusters using either k-means or LDA and then build a separate timeseries for each cluster. Then each timeseries is processed separately to detect whether it is stable or non-stable. For this step we use a sequence-to-sequence neural network, which is trained to jointly predict non-stable trends and pivot points. For comparison, we use linear regression as a baseline.

4.1. Building Timeseries

Clustering We use doc2vec model [21] to obtain document representations. The inferred document vectors are then clustered using k-means. Clustering is run independently for each of the 1000 datasets, so each dataset simulates a single independent use case. We set the number of clusters to 20 for all our datasets. Thus, we do not use our prior knowledge about the number of categories used. Moreover, perfect clustering is not possible with this setting since the number of clusters is bigger than the number of categories used to generate a dataset. The rationale behind this is that when working with real data we would not know the number of discourses in the collection. The method we propose does not aim at perfect clustering, only on detection of non-stable trends.

Clustering is done jointly for all time points in the dataset. Then we built a timeline for each cluster, by counting the number of documents from each cluster at each time point. Timelines are scaled to [0,1] interval so that the biggest value for each timeline is always 1.

LDA We use topic modelling as an alternative to k-means. We train a separate LDA model for each synthetic dataset and train with 20 topics to align with k-means.

The timeline on top of LDA is built using soft clustering, since an article can have more than one topic. To count the number of documents that belong to a certain topic, we use all documents where the topic probability is higher than 0.25. If no topic has a probability above the threshold, we assign the document to the topic with the highest probability. Similar to k-means, topic timelines are scaled to [0, 1] range.

Training Data The cluster-based timeseries, described above, are used only to construct the validation set. To train a neural network, we directly sample the patterns with noise to mimic the sequence of frequency in the clustered set. Stable and non-stable timeseries are sampled equally for the training.

The input for our models is a sequence of frequencies. The model produces two outputs: a binary prediction of whether a timeseries is stable or non-stable and a sequence, where the value at each time point is the probability that the time point belongs to a non-stable pattern. In the training data, we set to 1 all values between pattern start and end, while all other values are set to 0. If the timeseries is stable, all values in the output sequence are zeros, which corresponds to zero value for the first output.

4.2. Sequence-to-sequence model

Recurrent Neural Network The model structure is presented in Figure 3. The input to this model is a matrix with the shape $(N, 100)$ where N is the batch size and 100 is the timeseries length. Each example is a sequence of numbers in the range $[0, 1]$.

We use an RNN variant—bidirectional Long Short Term Memory (bi-LSTM)— stacked with one fully connected (FC) layer. The bi-LSTM layer has 256 hidden units.

The following FC layer takes all outputs from the LSTM layer and flatten them as input. Dropout layer is introduced to reduce overfitting. The FC layer is connected to two output layers: one to predict the probability that the input is non-stable and the other to predict a sequence of non-stable point probabilities. Both output layers use the sigmoid activation function to get probability values.

Convolutional Neural Network The CNN is intended for capturing local features for image recognition [22]. Our idea is to use this ability to detect patterns in sequence data. The CNN model is shown in Figure 4. The input and output is the same as one described for RNN . Because our sequence data only has one dimension, the 1D CNN layers are used for feature extraction. We use two stacked convolutional layers with a kernel size of 3. The first layer has 8 output channels while the second one expands to 16 channels. We also have max pooling layers after each convolutional layer. Finally, the output features are flattened and passed to the FC layer, and the rest of the model is organized identically to the RNN model.

Combined Model While RNN is good at handling sequence information, CNN is more suitable for local pattern detection. We leverage the strengths of both models to produce a combined model that might be more robust at pivot point detection. The architecture of the combined model (which we further denote as RCNN) is presented in Figure 5. CNN and bi-LSTM layers are identical to those used in the separate models. Then the hidden state output of the bi-LSTM layer is concatenated with the output of the last convolutional layer, flattened, and passed to the FC layer. After concatenating RNN and CNN outputs, the rest of the model is organized identical to the previous cases.

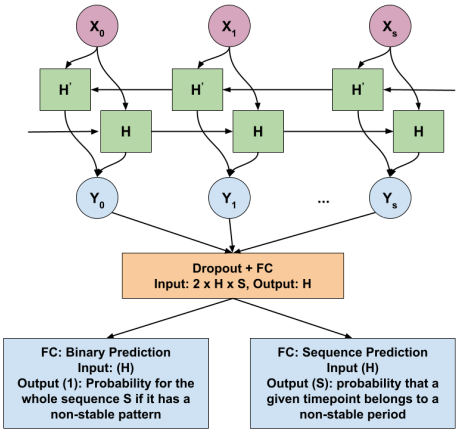


Figure 3: RNN network architecture with biLSTM layer. The prediction Y from all timesteps are used for the FC layer.

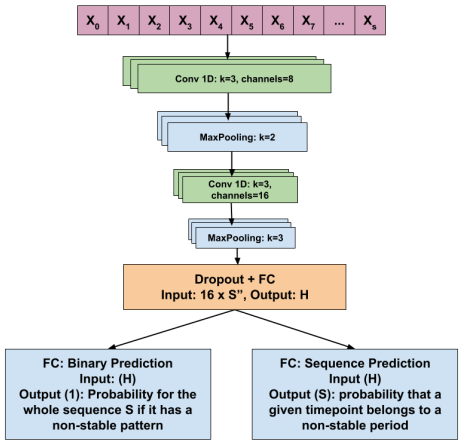


Figure 4: CNN network architecture. Where k is the kernel size, H is the hidden size, and S" is the length of sequence after going through the convolutional layers.

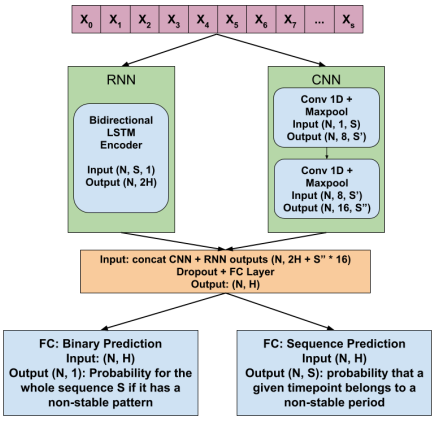


Figure 5: Combined (RCNN) network architecture. Where N is the batch size, H is the hidden size. S is the length of input sequence, S" is the length of CNN output. The hidden states from LSTM are used for the next layer instead of the predicted outputs.

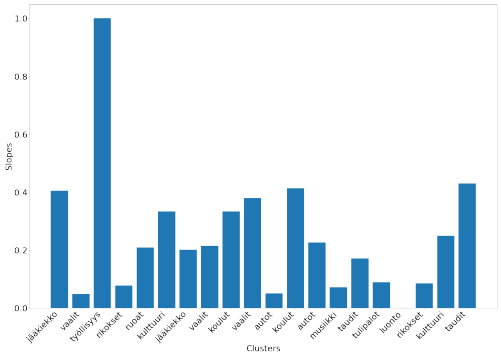


Figure 6: An example dataset, where each cluster, obtained from k-means, is fitted with linear regressions. The normalized slopes are shown in the histogram, with one pattern having significantly higher slope than the others, which indicates non-stable discourse dynamic. Bars are labelled with the major category of the articles within the cluster.

4.3. Baseline

Unlike neural models, our baseline is not independent for each cluster within a dataset. We fit a linear regression model to each of the 20 clusters obtained for the dataset. The absolute slope value of the linear function is normalized to a [0,1] scale, so that the largest normalized slope is

equal to 1. A timeseries with a slope above a certain threshold is then classified as non-stable. After preliminary experiments we set this threshold to 0.8 for all datasets.

As an example in Figure 6 we show an output for the dataset presented in Figure 1. In the histogram each bar is a cluster labeled with its major category, i.e. the most frequent category for the clustered articles. The y-axis is the normalized slope value. We see that the category for the biggest bar—*työllisyys*, employment—is the same as one used to build the increasing pattern in Figure 1.

Timeseries identified as non-stable in the previous step are processed using the sliding-window segmentation method to identify pivot points.[4]

5. Evaluation

Category-level Evaluation Category-level accuracy measures how well a model can detect a non-stable category. For each cluster classified as non-stable we define a *major category*, i.e. a category that has a highest count in this cluster. If this major category is the same as the target category used for the dataset generation, then prediction is considered to be correct. For each dataset we calculate a ratio of correct non-stable clusters to all non-stable clusters. If a model does not find any non-stable cluster for the dataset, the accuracy is 0.

Document level Precision, recall and F-measure are used to measure how ”clean” are subsets of documents that form non-stable patterns. For this evaluation, we use all clusters that are predicted to be non-stable, even if their major category is incorrect.

For each non-stable cluster, precision is calculated as a proportion of documents from the target category in this cluster, and recall as the proportion of documents from a non-stable cluster in a target category. The dataset recall and precision are the means of all non-stable cluster measures, and F-measure is computed as the harmonic mean of recall and precision. If all clusters are predicted to be stable then precision, recall and F-measure are set to zero. Then three measures are averaged across datasets.

Time-point level For each cluster that is classified as non-stable, a model must output time points where the non-stable pattern starts and ends. These pivot points segment a timeline into several periods. Then each pair of time points could belong either to the same or to different time periods. RandIndex [23] is computed as a proportion of time-point pairs correctly put either in the same or in the different periods. Shifting a pivot point by 1-2 positions from the true point slightly decreases RandIndex. Radical misplacement or finding an incorrect number of pivot points, however, results in a large drop.

RandIndex is averaged for all non-stable clusters in the dataset. If all clusters are classified as stable, RandIndex is zero. This measure is then averaged across all datasets. Note that this evaluation is orthogonal to the document-level measures, since it is possible to place pivot points to correct positions even if a cluster is noisy or incomplete.

[4]For more detail see the Rupture documentation: https://centre-borelli.github.io/ruptures-docs/user-guide/detection/window/

6. Experiments

Synthetic datasets Table 1 shows results obtained on the synthetic datasets with aforementioned measures. One of the most important results for us is the diversity of the model performance: this means that synthetic datasets are adequately complex and allows for method comparison.

The best performing model is the proposed combination of RNN and CNN (RCNN), which gives the highest results in combination with both k-means and LDA. The best performance is obtained by applying the combined model on top of the k-means output. On top of LDA the combination also yields the highest performance. CNN is better than RNN at non-stable pattern detection. However, RNN yields a much higher RandIndex, which means better at pivot point detection.

The lower performance of LDA compared to k-means needs to be investigated further. Obviously, LDA is much more than just a clustering technique: LDA is a Bayesian model, which outputs topic distribution over documents. In our experiments, this distribution is converted into hard labels and used only indirectly. It is likely that a higher performance could be achieved by other ways of combining topic modelling with neural networks. There is another difficulty when it comes to a rich morphological language like Finnish, where words have many variants and compounds are frequently used [24].

Experiments with real data For a qualitative assessment, we use another Finnish corpus: The Finnish News Agency (STT) Archive [5]. We limit our experiments to the data from years 2007-2008, so does not overlap in time with the YLE dataset. We split the two year data into weeks, excluding the first and the last two weeks, which gives us 100 weeks. Then we can directly apply models trained on synthetic data. The dataset consists of 250,000 documents.

We use our best model for this experiment, i.e. combined RCNN applied on top of k-means with 20 clusters. Out of those 20 clusters, 6 were classified as unstable. We manually scanned the documents within these clusters and found a couple of clusters for which we could find an interpretation. For example, Figure 7 shows a cluster associated with party politics. The date of the Finish parliamentary elections, shown with the green vertical line, is positioned between two automatically determined pivot points—it seems natural that elections are actively discussed some time before and after the event.

We use this experiment to demonstrate that a model trained on synthetic datasets, generated using the proposed procedure, is able to extract meaningful results from real-world data. Whether these results would be relevant for digital humanities or computational social science research is yet to be found in collaboration with domain specialists. Previous collaborations[10] indicate that there is a need for a model able to track discourse change in textual data.

Conclusion

We presented the novel task of automatic detection of discourse change in text collections, which is relevant for historical research and digital humanities in general. However, computational

[5] http://urn.fi/urn:nbn:fi:lb-2019041501

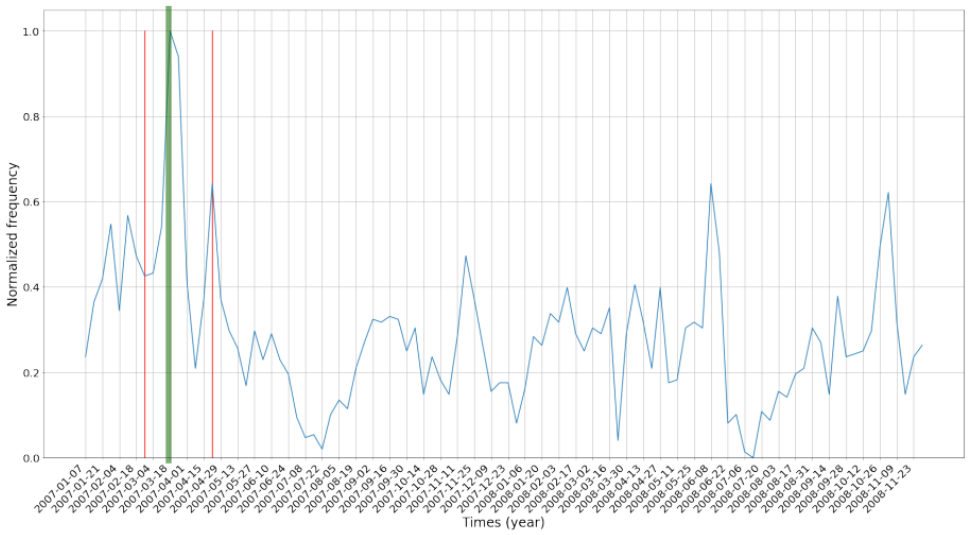


Figure 7: A non-stable cluster obtained on the STT data. Red lines: detected pivot points. Green line: the Finnish Parliamentary elections.

Method		Category accuracy	Precision	Recall	F1	Rand index
STEP 1	STEP2					
k-means	Regression	52.78	43.98	34.73	37.04	42.52
	RNN	73.63	60.55	46.33	50.43	73.17
	CNN	75.17	61.46	46.56	51.49	67.79
	RCNN	**78.43**	**63.77**	**51.69**	**55.22**	**73.26**
LDA	Regression	41.88	31.56	31.26	27.14	41.04
	RNN	38.65	30.48	31.84	27.53	65.04
	CNN	47.73	36.41	33.26	31.87	53.27
	RCNN	51.46	37.22	43.94	36.03	60.43

Table 1: Result obtained on 1000 synthetic datasets

methods to tackle this type of problems are not yet established. One of the main obstacles is the lack of training data and fundamental difficulty to annotate corpus-level phenomena. To overcome this issue we proposed a methodological framework that leans to discourse-change simulation with synthetic data, which allows us to train supervised models. The procedure which we proposed in this paper generates sufficiently complex datasets so that the problem cannot be solved by simple methods, such as regression. This allows for evaluation, comparison, and improvement of the methods, impossible on most typical use cases where ground truth is not accessible.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable comments regarding the positioning of this work and historical discourse studies in general. Though we were unable to address these issues in this paper we take them seriously and will take into consideration in our future work.

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

References

[1] L. Viola, J. Verheul, Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920, Digital Scholarship in the Humanities 35 (2020) 921–943.

[2] R. Light, J. Cunningham, Oracles of peace: Topic modeling, cultural opportunity, and the Nobel peace prize, 1902–2012, Mobilization: An International Quarterly 21 (2016) 43–64.

[3] T.-I. Yang, A. Torget, R. Mihalcea, Topic modeling on historical newspapers, in: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011, pp. 96–104.

[4] D. J. Newman, S. Block, Probabilistic topic decomposition of an eighteenth-century American newspaper, Journal of the American Society for Information Science and Technology 57 (2006) 753–767.

[5] S. Hengchen, R. Ros, J. Marjanen, A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950, in: Proceedings of the Digital Humanities (DH) conference, 2019.

[6] M. Kestemont, F. Karsdorp, M. During, Mining the twentieth century's history from the Time magazine corpus, in: Abstract book of EACL 2014: the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, p. 62.

[7] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 113–120.

[8] A. B. Dieng, F. J. Ruiz, D. M. Blei, The dynamic embedded topic model, arXiv preprint arXiv:1907.05545 (2019).

[9] X. Wang, A. McCallum, Topics over time: a non-markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 424–433.

[10] J. Marjanen, E. Zosa, S. Hengchen, L. Pivovarova, M. Tolonen, Topic modelling discourse dynamics in historical newspapers (2021).

[11] D. Hall, D. Jurafsky, C. D. Manning, Studying the history of ideas using topic models, in: Proceedings of the 2008 conference on empirical methods in natural language processing, 2008, pp. 363–371.

[12] N. Tahmasebi, L. Borin, A. Jatowt, Survey of computational approaches to lexical semantic change, in: Preprint at ArXiv 2018., 2018. URL: https://arxiv.org/abs/1811.06278.

[13] A. Kutuzov, L. Øvrelid, T. Szymanski, E. Velldal, Diachronic word embeddings and semantic

shifts: a survey, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1384–1397.

[14] X. Tang, A state-of-the-art of semantic change computation, Natural Language Engineering 24 (2018) 649–676.

[15] D. Schlechtweg, S. S. im Walde, S. Eckmann, Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 169–174.

[16] A. Tsakalidis, M. Liakata, Sequential modelling of the evolution of word representations for semantic change detection, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8485–8497.

[17] D. Schlechtweg, S. S. i. Walde, Simulating lexical semantic change from sense-annotated data, in: The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)., 2020.

[18] P. Shoemark, F. F. Liza, D. Nguyen, S. Hale, B. McGillivray, Room to glo: A systematic comparison of semantic change detection approaches with word embeddings, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 66–76.

[19] A. Rosenfeld, K. Erk, Deep neural models of semantic shift, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 474–484.

[20] R. A. Blythe, W. Croft, S-curves and the mechanisms of propagation in language change, Language (2012) 269–304.

[21] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, 2014, pp. 1188–1196.

[22] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117. URL: http://dx.doi.org/10.1016/j.neunet.2014.09.003. doi:10.1016/j.neunet.2014.09.003.

[23] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods, Signal Processing 167 (2020) 107299.

[24] Q. Duong, M. Hämäläinen, S. Hengchen, An unsupervised method for ocr post-correction and spelling normalisation for finnish, 2020. arXiv:2011.03502.

Appendix H: Topic modelling discourse dynamics in historical newspapers

Topic Modelling Discourse Dynamics in Historical Newspapers

Jani Marjanen[1*][0000−0002−3085−4862], Elaine Zosa[2*][0000−0003−2482−0663], Simon Hengchen[3][0000−0002−8453−7221], Lidia Pivovarova[2][0000−0002−0026−9902], and Mikko Tolonen[1][0000−0003−2892−8911]

[1] Helsinki Computational History Group, University of Helsinki
[2] Department of Computer Science, University of Helsinki
[3] Språkbanken, University of Gothenburg[‡]
firstname.lastname@{helsinki.fi,gu.se}

Abstract. This paper addresses methodological issues in diachronic data analysis for historical research. We apply two families of topic models (LDA and DTM) on a relatively large set of historical newspapers, with the aim of capturing and understanding discourse dynamics. Our case study focuses on newspapers and periodicals published in Finland between 1854 and 1917, but our method can easily be transposed to any diachronic data. Our main contributions are a) a combined sampling, training and inference procedure for applying topic models to huge and imbalanced diachronic text collections; b) a discussion on the differences between two topic models for this type of data; c) quantifying topic prominence for a period and thus a generalization of document-wise topic assignment to a discourse level; and d) a discussion of the role of humanistic interpretation with regard to analysing discourse dynamics through topic models.

Keywords: Discourse Dynamics, Finland, Historical Newspapers, Nineteenth Century, Topic Modelling.

1 Introduction

This paper reports our experience on studying discursive change in Finnish newspapers from the second half of the nineteenth century. We are interested in grasping broad societal topics, discourses that cannot be reduced to mere words, isolated events or particular people. Our long-lasting goal is to investigate a global change in the presence of such topics and especially finding discourses that have disappeared or declined and thus could easily slip away in modern research. We believe that these research questions are better approached in a data-driven way without deciding what we are looking for beforehand, though the choice of the most suitable techniques for such research is still an open problem.

In this paper we focus on developing methodology. Choosing available algorithms for analysis guides possible outcomes as they are designed to be operationalised in

[‡]SH was affiliated with the University of Helsinki for most of this work.
[*]Equal contribution.

certain ways. Approaching our goal with mere word counts is counterproductive due to the sparseness of the language and the variety of discourse realisations in a given text. Further, word counts are unreliable with historical data due to never ending language change, spelling variations and text recognition errors.

Thus, as many other papers in the area of digital humanities, we utilize topic modelling as a proxy to discourses. In particular, we apply the "standard" Latent Dirichlet Allocation model [3, LDA] and its extension the Dynamic Topic Model [2, DTM], which is developed specifically to tackle temporal dynamics in data. However, any model has its limitations and tends to exaggerate certain phenomena while missing other ones. We focus on the difference between models and try to reveal their limitations in historical data analysis from the point of view that is relevant for historical scholarship.

Our main contributions are the following:

- We propose a **combined sampling, training and inference procedure** for applying topic models to large and imbalanced diachronic text collections.
- We discuss differences between two topic models, paying special attention to how they **can be used to trace discourse dynamics**.
- We propose a method to quantify **topic prominence for a period** and thus to generalize document-wise topic assignment to a discourse level.
- We **acknowledge and discuss the drawbacks of topic stretching,** which is typical for DTM. It is commonly known that DTM sometimes represents topics beyond the time period, but thus far there is no discussion in how researchers should tackle this for humanities questions.

In order to illustrate the appropriateness of the proposed methodology we discuss two use cases, one relating to discourses on church and religion and one that relates to education. The role of religion and education has been studied extensively in historical scholarship but there are no studies that deal with these topics through text mining of large-scale historical data. These two topics were chosen due to the the fact that the former was in general a discourse in decline relating to the process of secularization in Finnish society, whereas the latter increased in the second half of the nineteenth century and relates to the modernization of Finnish society and the inclusion of a larger share of the population in the sphere of basic education. In addition to these two interlinked discursive trends, we also use other examples to illustrate the strengths and weaknesses of LDA and DTM for this type of historical research.

2 Data

Our dataset is from the digitised newspaper collection of the National Library of Finland (NLF). This dataset contains articles from *all* newspapers and most periodicals that have been published in Finland from 1771 to 1917. Several studies have used parts of this dataset to investigate such issues as the development of the public sphere in Finland, the evolution of ideological terms in nineteenth-century Finland and the changing vocabulary of Finnish newspapers [36, 17, 16, 11, 21, 22, 25, 29, 12].

The full collection includes articles in Finnish, Swedish, Russian, and German. In this work we focus only on the Finnish portion starting from 1854 because this is the point where we determined we have sufficient yearly data to train topic models. The resulting subset has over 3.6 million articles and is composed of over 2.2 billion tokens. Figure 1a shows that the number of tokens published per year in Finnish-language papers increased steadily. The average article has 526 tokens but article length varies widely from year to year, as seen in Figures 1b and 1c which show the average article length and the number of articles per year. As made clear by these figures, there is a noticeable difference in the number of articles and average article length after 1910. This shift does not reflect the actual articles in the newspapers, but is the result of a change of OCR engine used to digitise the collection [20]. While the raw data is publicly available, we used the lemmatised version of the newspaper archive produced by Eetu Mäkelä, whom we thank.

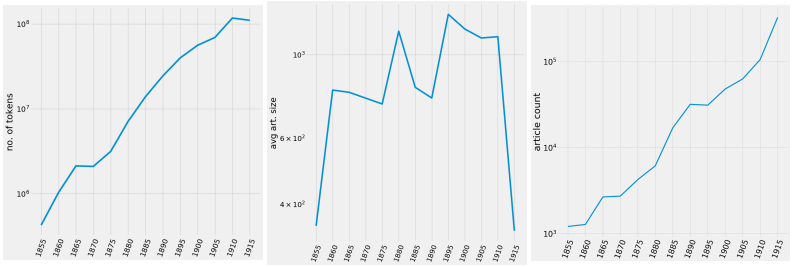
Still, even if the article segmentation differs in the latter period, Fig. 1a shows that there is steady increase in the vocabulary used in the Finnish-language newspapers published in the second half of the nineteenth century. They also covered more themes and regions. This entailed a process of diversification and modernization of the Finnish press, which has been widely discussed in historiography. As a collection, the newspapers vary a lot in style and focus. Some larger newspapers mainly contain political content, whereas others are rather specialised, and yet others thrived by giving a voice to the local public [35, 22, 16, 32]. This means that any analysis done on the entirety of the newspapers, like topic models, tend to balance out some of the differences between newspapers. This variety in the content, is also something that make newspapers such an interesting source material for historical research that is interesting in an overview of society. Although some issues were obviously not discussed because of taboo, courtesy or censorship, most of the themes present in public discourse are recorded in the newspapers and thus accessible to us in the present. Hence, we believe newspapers are an especially good source of assessing how the role of particular discourses changed over time.

2.1 Preprocessing the data

Given the size of the data and its inherent nature, notoriously the OCR quality and the unbalanced data from different time slices, we performed a series of pre-processing steps on the data.[1]

Despite prior work (albeit on English), showing that stemming has no real advantage for likelihood and topic coherence and can actually degrade topic stability [30], we follow [40, 10, 13] and use a lemmatised version of the corpus. Indeed, the work in [10] hints at the fact that Finnish, being much more inflected than English, would benefit from lemmatisation, whereas in [40, 13] the authors stem so as to reduce the huge number of token types due to OCR issues which impacts the performance of topic

[1]The more apt phrase "purposeful data modification", coined by [34], advocates that our material is not mere data that can go through a standardised "pre-processing" pipeline. Rather, the data is modified and altered only for the specific purposes of this study, and following this study's technical and scientific requirements only.



(a) Corpus size (b) Average article size (c) No. of articles

Fig. 1: Characteristics of the NLF dataset

modelling [38]. After lemmatisation, we remove tokens that occur less than 40 times in the collection, stopwords, punctuation marks and tokens with less than 3 characters. These are additional measures to further reduce the vocabulary size and mitigate the impact of OCR noise.

3 Topic Models

3.1 LDA

Topic modelling is an unsupervised method to extract topics from a collection of documents. Typically, a topic is a probability-weighted list of words that together express a theme or idea of what the topic is about. One of the most popular topic modelling methods currently in use is Latent Dirichlet Allocation (LDA), which is "a generative probabilistic model for collections of discrete data such as text corpora" [3]. It has been extensively used in the digital humanities to extract certain themes from a collection of texts [4]. In this model, a document is a mixture of topics and a topic is a probability distribution over a vocabulary. A limitation of LDA for historical research, in its vanilla form, is that it does not account for the temporal aspect of the data: every document in the collection is "considered synchronic", as time is simply not a variable in the model. Many document collections such as news archives, however, are diachronic—the documents are from different points in time, and scholars wish to study the evolution of topics.

There are different ways to overcome this limitation. One possibility is to split the data into time slices and train LDA separately on each slice. However, in this case LDA models for each slice would be independent of each other and there is no straightforward approach of matching topics from independent models trained on disjoint data. Another possibility, which we explore in this paper, is to train a single model for a subset of the whole data set over the entire time period and then use *topic prominence* as proxy for the dynamics of discourses over time.

To do this, we compute the prominence of a topic in a given year by summing up the topic contribution for each document in that year and then normalise this number by the sum of all topic contributions from all topics for that year, as in Equation 1.

$$P(z_k|y) = \frac{\sum_{j=1}^{|D_y|} P(z_k|d_j)}{\sum_{i=1}^{T}\sum_{j=1}^{|D_y|} P(z_i|d_j)} \quad (1)$$

where y is a year in the dataset, k is a topic index, D_y is the number of documents in year y, d_j is the j^{th} document in year y and T is the number of topics in the model.

The large size of the collection and its unbalanced nature is a problem for training topic models. It is computationally expensive to train a model with millions of articles and the resulting model would be heavily biased towards the latter years of newspaper collection because it has far more data. To overcome these issues, we sampled the collection such that we have a roughly similar data size for each year of the collection and as a result, we also get a vastly reduced dataset. However, to have a model of discourse dynamics that reflects the collection more closely, we compute topic prominence using the entire collection and not just the sampled portion. We do this by inferring the topic proportions of all the documents in the collection and using these inferred distributions to compute topic prominence.

3.2 DTM

As mentioned above, there are topic models that explicitly take into account the temporal dynamics of the data. One such model is the dynamic topic model (DTM). DTM is an extension of LDA that is designed to capture dynamic co-occurence patterns in diachronic data. In this model, the document collection is divided into discrete time slices and the model learns topics in each time slice with a contribution from the previous time slice. This results in topics that evolve slightly–words changing in saliency in relation to a topic–from one time step to the next.

However, DTM also has its own limitations. It is based on an assumption that each topic should be to some extent present in each time slice, which is not always the case with real-world data such as news archives where events and themes can sometimes disappear and then re-appear at some point in the future.

Perhaps more importantly for historical research, a weakness of DTM lies in its design: to accomplish alignment across time the topic model is fit across the whole vocabulary and thus smoothing between time slices is applied. As a result, events end up being "spread out" before and after they are known to happen. This problem only becomes evident after a thorough analysis: similar models in different fields such as lexical semantic change present the same issue – the dynamic topic model SCAN [7] generates a "plane" top word for the year 1700 (two centuries ahead of the Wright Flyer, and well before the word's first attested sense of "aeroplane"), while similar model GASC [26, 23] encounters the same weakness when modelling Ancient Greek. There is unfortunately no easy way to bypass this obstacle, which is particularly problematic when studying historical themes.

For both the LDA and DTM models, we use the Gensim implementation [28] with default model hyperparameters.

4 Related Work

Topic models are widely used in the digital humanities and social sciences to draw insights from large-scale collections [4] ranging from newspaper archives to academic journals. In this section, which we do not claim to be exhaustive, we discuss some of the previous works that aimed to capture historical trends in large data collections or used such collections to study discourses using topic models. All in all, these examples highlight that there is a need to discuss how topic models can be used to capture discursive change.

In [24] the authors use Latent Semantic Analysis, another topic modelling method, to study historical trends in eighteenth-century colonial America with articles from the *Pennsylvania Gazette*. Their work also used topic prominence to show, for instance, an increased interest in political issues as the country was heading towards revolution. The authors of [40] fit several topic models on Texan newspapers from 1829 to 2008. To discover interesting historical trends, the authors slice their data into four time bins, each corresponding to historically relevant periods. Such a slicing is also carried out in [9], where the author fits LDA models on Dutch-language Belgian socialist newspapers for three time slices that are historically relevant to the evolution of workers rights, with the aim of generating candidates for lexical semantic change.

Topic modelling has also been used in discourse analysis of newspaper data. In [37] the authors applied LDA to a selection of Italian ethnic newspapers published in the United States from 1898 to 1920 to examine the changing discourse around the Italian immigrant community, as told by the immigrants themselves, over time. They proposed a methodology combining topic modelling with close reading called discourse-driven topic modelling (DDTM). Another study examined anti-modern discourse in Europe from a collection of French-language newspapers [5]. In this case, however, the authors primarily use LDA as a tool to construct a sub-corpus of relevant articles that was then used for further analysis. Modernization was also an issue in the study of Indukaev [14], who uses LDA and word embeddings to study changing ideas of technology and modernization in Russian newspapers during the Medvedev and Putin presidencies.

LDA was not designed for capturing trends in diachronic data and so several methods have been developed to address this, such as DTM, Topics over Time [39, TOT], and the more recent Dynamic Embedded Topic Model [6, DETM], an extension of DTM that incorporates information from word embeddings during training. As far as we are aware, DTM and TOT have not been used for historical discourse analysis or applied to large-scale data collections. In the original papers presenting these methods, DTM was applied to 30,000 articles from the journal *Science* covering 120 years and TOT was applied to 208 State of the Union Presidential addresses covering more than 200 years. This was to demonstrate the evolution of scientific trends for the former and the localisation of significant historical events for the latter. Recently DETM was applied on a dataset of modern news articles about the COVID-19 pandemic where the authors observed differences between countries in how the pandemic and the reactions to it were framed [19].

In the mentioned cases researchers tackle the interpretative part of using topic models for humanistic research in different ways. Like Pääkkönen and Ylikoski [27] state, they toggle between some sort of topic realism, that is, using topic models to grasp

something that exists in the data, and topic instrumentalism, that is, using topic models to find something that can be further studied. Only Bunout [5] is a clear case of topic instrumentalism. All the other studies depart from some sort of realist position, and attempt to grasp policy shifts, ideas, discourses or framings of topics through topic models, but end up with correctives of some kind by highlighting the interpretative element [24, 37], by deploying formal evaluation by historians [9] or by using other quantitative methods to fine tune the results [14]. The interpretative aspect seems especially important when it comes to deciding on what researchers use the topics to study as they can reasonably relate to historical discourses, the semantics of related words, or simply ideas. How the topics are seen to represent these or, more likely, how the researchers use the topics to make an interpretation about these based on the topics, requires a strong element of interpretation [27]. Studies show that interpreters prefer to be able to go back to actual texts in order to make sense of topics [18], which is more than reasonable, but it also seems that there is a further need for researchers to understand how different topic-modelling methods represent diachronic data. Without this knowledge it is difficult to assess to which degree and for which time periods researchers need to manually assess individual documents.

5 Use Cases

What a discourse is, has been heavily theorised within the different strands of discourse analysis [1], but the advent of digital methods that can handle large textual data sets require quite some adjustment of discourse analysis as we know it. Like this article, others have turned to topic models to grasp changes in discourse [37, 5], but this article seeks specifically to discuss the interpretation that is required when we use topic models to study discourse dynamics. The probabilistic topic models set clear boundaries between topics and in doing so might merge or separate things that historians might regard as coherent topics. However, where the probabilistic model enforces boundaries, human interpretation in general is very bad at setting those boundaries and usually just identifies the core of a discourse or topic, but cannot say where it ends.

To get at the tension between topics and discourses, we approached the material without a predefined idea about which topics we wanted to study in order to keep the study as data-driven as possible. Our interest was to use topic modelling to capture topics that could in a meaningful way be related to societal discourses, that is themes that cannot be narrowed down to individual words, but still are reasonably coherent and form at least loose topics. To this end, we trained topic models with $k \in \{30; 50\}$, inferred topic distributions for the whole collection and inspected models by carefully going through the top words in each topic and using PyLDAVis[2] [31] to study overlap between topics and salience of terms per topic in LDA and heatmap visualizations for DTM. All topics were annotated and evaluated from the point of view of historical interpretation. We then opted to use the 50-topic model to study discourse changes over time. As is common, a portion of the topics seemed incoherent or were clearly the result of the layout in newspapers (e.g. boilerplate articles about prices etc.) and

[2] https://github.com/bmabey/pyLDAvis

did not produce interesting information about societal discourses. Further, some of the topics clearly overlap, so that a cluster of 2-5 topics can reasonably be seen as related to a particular societal discourse. The advantage of choosing 50 topics over 30 lies precisely in the possibility of merging topics later on in interpretation, while splitting them is more difficult.

To discuss the benefits of LDA and DTM, we chose to focus on two specific themes, the discourse relating to religion and religious offices, and education. They are both rather neatly identifiable in the data, but display different trends. The former is in decline over the period of interest, whereas the latter increases in topic prominence. They can also be related to large scale processes in Finland, religious discourse to the secularization of society and education to the modernization of civic engagement.

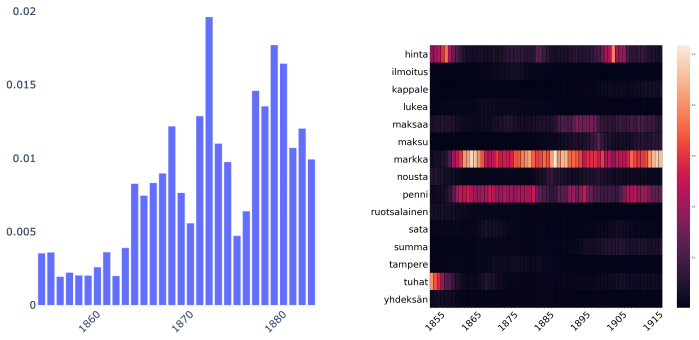
5.1 DTM and Stretching of Topics

The two topic modelling methods perform in somewhat different ways. As mentioned, DTM is designed to incorporate temporal change in the topics, which means it includes a stronger sense of continuity in its representations of data. Whether or not this is desirable, depends on the research question, but our contention is that for studies interested in discursive change, this is either a problem or at least it is something that needs to be factored in making the historical interpretation. If we want to understand when certain discourses became dominant, declined, or even disappeared, this type of stretching cannot be allowed.

An exceptionally illustrative example of stretching among our fifty topics, is an introduction of the Finnish mark as a currency (Fig. 2a). With top words such as “mark”, “penny”, “price”, “thousand”, “pay” etc. the topic comes across as one with high internal coherence. We also see that the topic grows in prominence over time, from being relatively modest in the 1850s to gradually increased prominence after 1860. This makes sense, as the mark was adopted as currency in the year 1860 and after that self-evidently figured in public discourse. However, when we look at a heatmap visualization of the topic (Fig. 2b), we see how the topic stretches from the period 1854–1859 to the period 1860–1917, that is, from the period before the introduction of the mark to the period it was in use. After 1860 the words “mark” and “penny” are by far the most dominant terms in the topic, but for the period before 1860, the dominant terms are “price” and “thousand.” It is clear that “mark”, “penny”, “price”, and “thousand” are words that can belong to the same topic, but the heatmap representation clearly shows that the focus in the topic shifts. It is almost as if two related topics are merged as to represent one topic over the whole time period. In a situation where a historical interpretation highlights a change in past discourse, DTM produces continuity.

While there is obviously no right answer as to when one topic is stretched a bit or when different topics are simply merged together to provide a temporally continuous topic, it seems that DTM is especially problematic if one wants to study discourses that emerge or disappear in the middle of a time period studied. This means that any historical analysis using DTM requires a component of historical interpretation of not only topic coherence, but also topic coherence *over time*. Here, relying on word embeddings like in [14] can help, but this is primarily a task for evaluating the topics.

71



(a) Introduction of the Finnish mark in 1860 (y-axis indicates the topic probability)

(b) Heat map of terms linked to the introduction of the Finnish mark in 1860.

Fig. 2: Topic related to the introduction of the Finnish mark in 1860 (DTM). The most prominent terms in the heatmap are are "Mark" = *markka*, "penny" = *penni*, "price" = *hinta*, "thousand" = *tuhat*, "pay" = *maksu* and *maksaa*.

The speed of topic evolution can be controlled by a parameter in the DTM model. However, the 'ideal' amount of stretching is difficult to assess. For analysing discourse, this might in some cases be productive as it can point at links between nearby discourses, but is largely problematic as it hides discontinuities in the data. It becomes even problematic when dealing with material factors, like the introduction of the Finnish mark, as the stretching effect is likely to produce anachronistic representations, that is, placing something in the wrong period of time. Dealing with anachronism can perhaps be seen as one of the cornerstones of the historian's profession, which makes DTM as an anachronism prone method a poor match for historical study. Avoiding anachronisms completely is impossible, most historians would agree, but knowing when to avoid them and how to communicate about anachronistic elements in historical interpretation is key to history as a discipline [33].

5.2 Religion and Secularization

Our model performed well in grasping topics that relate to religion. The initial expectation regarding the discourse dynamics was that religious topics would be in decline. We hoped that using a topic model would be a way of showing this quantitatively. Results obtained from both LDA and DTM, presented in Figures 3a and 3b respectively, harmonize with our initial hypothesis, but do so differently. The DTM and LDA outputs cannot be aligned in any other way than manual interpretation by domain experts. In doing this we simply regarded topics that included several words that denote religious practices or offices as religious. Thus, the definition of "religious" is is rather narrow, but it also seems to match the topics that emerged from our data.

In order to inspect the discourse dynamics of religious topics, we have combined several topics that related to religious themes in the LDA model, whereas in the latter, DTM model, we only chose one topic to be represented.[3]

To our knowledge, topic models have not been used to study discursive change regarding secularization. However, in line with some earlier qualitative assessments [15], we hypothesize that this decline in religious discourse entails two interrelated developments: 1) Religion did not disappear from public discourse, but instead changed and disappeared from certain *types* of discourses. In the early nineteenth century, religion had a much more holistic presence in public discourse, meaning that religious metaphors and religious expressions and topics were used at a much vaster scale. 2) Over the course of the nineteenth century, religious topics became more focused. This means a segmentation of public discourse so that religious topics were increasingly confined to particular journals or genres.

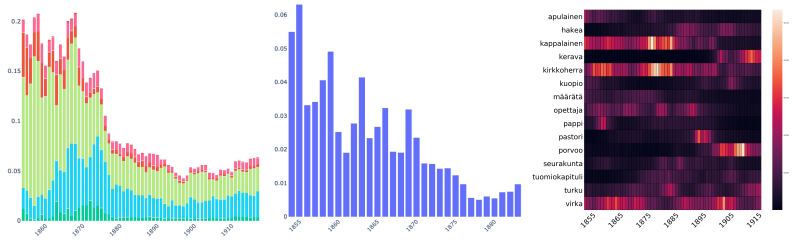
Keeping in mind the issue of stretching with DTM, we can look into the shifting saliency of words within the topic of religious offices and notice a shifting focus over time (Fig. 3c). In the early 1900s terms relating to "holding an office" and names of particular congregations become more dominant in the topic. This, again, suggests that DTM as a method does some stretching. There is a downside and an upside to this. On the one hand, the stretching distorts the topic prominence a bit by making it look like there is more continuity than in the LDA visualization. However, this may not be that crucial as the declining trends in Fig. 3a and Fig. 3b are rather similar. On the other hand, the stretching may be good for detecting conceptual links between different groups of words. In this particular case the stronger link between religious offices and some towns like Kerava and Porvoo, is probably indicative of a move of religious discourse from an overarching question to something that is more likely dealt with in conjunction to matters at local parishes. That is, religious offices were more often than before dealt with in connection to local congregations. This is in line with our above-mentioned assumption about religious discourse becoming more distinct.

5.3 Education and Modernity

While we expected religious themes to decline and become less central, we assumed there would be some themes that partly overlap with religion, but also would show an increasing trend. One example of this is the topic of education, which has historically been heavily interwoven with the church, but at the same time when basic education became available for a higher amount of people, it also became central in questioning the role of the church and religion. Education in nineteenth-century Finland was both central for ensuring conformity of the Lutheran faith, but paradoxically also was a vehicle of secularization. [8]

As in the case of religious discourse, alignment between DTM and LDA can only be made through human interpretation. It seems, that in this case DTM captures one topic

[3]We also experimented with more data-driven methods to cluster topics, including for example methods based on Jensen-Shannon Divergence. They unfortunately did not lead to clusters that our domain experts would make sense of. Nonetheless, despite this, we still believe this is an interesting avenue to pursue which could help answer the common 'number of topics' question often brought up within the field.



(a) Topics related to religion on decline (LDA)

(b) Development of religious topic (chaplain, priest and office) over time

(c) Heatmap of terms linked to office of religion topic.

Fig. 3: Religious topics in LDA (a) and DTM (b,c); y-axis in (a, b) indicates the topics' probabilities. Most prominent terms in the heatmap are "chaplain" = *kappalainen*, "vicar" = *kirkkoherra*, "teacher" = *opettaja*, "priest" = *pappi*, "Porvoo" (a town), "parish" = *seurakunta*, "Turku" (a town), and "office" = *virka*.

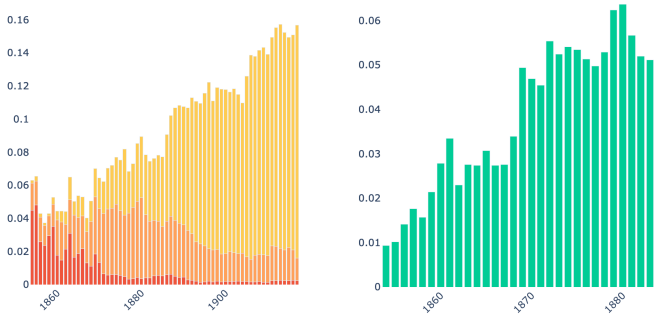
that is fairly coherent, revolves around education and schooling, and is on the rise in the research period (Fig. 4b). For LDA, this is not the case, as an PyLDAVis inspection of most salient words across all fifty topics show that words like "school" and "folk school" appear mostly in three topics of which two are in decline and one heavily on the rise (Fig. 4a).

Interestingly, LDA and DTM seem to be pointing at a similar historical development. The two declining LDA topics are based on their most salient terms and are more focused on schools as buildings and institutions as well as teaching as a profession, whereas the topic on the rise includes salient vocabulary relating to, not only schools, but also meetings, civic engagements, and decision making. The DTM topic at hand shows a similar development which can be inspected in a heatmap of most salient terms over time. The terms "school", "child", and "teacher" dominate early in the period. By the end of the period the topic becomes broader, and terms like "municipality" and "meeting" have become more salient than the vocabulary relating to schools. Here the stretching of DTM creates the links that are also visible in the three LDA topics, and it shows a transformation in which educational issues are present in the whole topic, but focus shifts from concrete schools to civic engagement.

6 Conclusions

Our focus in this text has been on discourses that cannot be reduced to mere words, isolated events or particular people, but concern broader societal topics that either declined or gained in prominence. The interpretation of these topics and their contextualisation to nineteenth-century Finnish newspapers revealed clear topical cores that can be interpreted as an encouraging point of departure for further explorations based on topic models when aiming to understand Finnish public discourse through historical newspapers.

74



(a) Development of education topic over time (LDA)

(b) Development of education topic over time (DTM)

Fig. 4: Education topic in LDA and DTM; y-axis indicates the topics' probabilities

In this paper, we have learned that although it is difficult to pinpoint exactly where a discourse or topic ends, LDA and DTM can fairly reliably grasp many semi-coherent themes in past discourse and help us study the dynamics of discourses. However, our comparison of LDA and DTM as methods for getting at past discourse also shows that both methods require a very strong interpretative element in analysing historical discourses. DTM is much more prone to stretch or even merge topics, which requires an interpretative assessment of whether the stretching highlights interesting historical continuities or if it hides historical discontinuities that would require attention. We found that producing heatmaps of term saliency over time for each topic is a very useful way of doing this type of assessment. For LDA, stretching is not so much a problem, but often it seems interpretation is needed in seeing which topics logically relate to one another. While historical discourse analysis is traditionally tied strongly to a tradition of hermeneutic interpretation, the use of topic models to grasp discourse dynamics does not remove that need even if they allow for a quantification of discourse dynamics over time.

While we regard stretching in DTM as a predominantly negative feature, in some cases it can be useful. In the topics relating to education discussed above, the stretching in DTM actually points out links in discourses and is quite productive for the interpretative process of trying to figure out discourse dynamics. However, also in this case, the relevance of historical interpretation should be highlighted because it is very hard to tell whether the stretching of topics is an accurate reflection of the data or a shortcoming of the model. This can be addressed only by relating visualisations of topics to existing historical research and reading source texts. Humanities scholars are in general very good at making such interpretations, but it also needs to be noted that when we move further into the domain interpretative scholarship, we also lose some of the benefits of working with quantifying models. While it would be foolish to claim that a topic model represents data in a way that it provides simple facts about historical development, our use cases show that if we seek to find more reliable quantification LDA may

provide better results than DTM. Further, using LDA moves the interpretative stage further down in the research process, as it is likely to be about evaluating the connections between different topics over time. In DTM, the interpretation is likely moved forward to an evaluation of how well the algorithm did this merging topics. On this sense, our take on topic models harmonises with [27] who stress the role of humanistic interpretation, but for the sake of transparency suggest pushing the interpretation stage later in the research process.

Acknowledgements

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA). SH is funded by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184).

References

1. Angermuller, J., Maingueneau, D., Wodak, R. (eds.): The discourse studies reader: Main currents in theory and analysis. John Benjamins Publishing, Amsterdam, the Netherlands ; Philadelphia PA (2014)
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine Learning. pp. 113–120 (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
4. Brauer, R., Fridlund, M.: Historicizing topic models, a distant reading of topic modeling texts within historical studies. In: International Conference on Cultural Research in the context of Digital Humanities, St. Petersburg: Russian State Herzen University (2013)
5. Bunout, E.: Grasping the anti-modern discourse on Europe in the digitised press or can text mining help identify an ambiguous discourse? (2020)
6. Dieng, A.B., Ruiz, F.J., Blei, D.M.: The dynamic embedded topic model. arXiv preprint arXiv:1907.05545 (2019)
7. Frermann, L., Lapata, M.: A Bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics **4**, 31–45 (2016)
8. Hanska, J., Vainio-Korhonen, K. (eds.): Huoneentaulun maailma: kasvatus ja koulutus Suomessa keskiajalta 1860-luvulle. Suomalaisen Kirjallisuuden Seuran toimituksia, 1266:1, Suomalaisen kirjallisuuden seura, Helsinki (2010), publication Title: Huoneentaulun maailma : kasvatus ja koulutus Suomessa keskiajalta 1860-luvulle
9. Hengchen, S.: When Does it Mean? Detecting Semantic Change in Historical Texts. Ph.D. thesis, Université libre de Bruxelles (2017)
10. Hengchen, S., Kanner, A.O., Marjanen, J.P., Mäkelä, E.: Comparing topic model stability between Finnish, Swedish, English and French. In: Digital Humanities in the Nordic Countries (2018)
11. Hengchen, S., Ros, R., Marjanen, J.: A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In: Proceedings of the Digital Humanities (DH) conference (2019)
12. Hengchen, S., Ros, R., Marjanen, J., Tolonen, M.: A data-driven approach to studying changing vocabularies in historical newspaper collections. Digital Scholarship in the Humanities (2021)

13. Hill, M.J., Hengchen, S.: Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. Digital Scholarship in the Humanities **34**(4), 825–843 (2019)
14. Indukaev, A.: Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings. In: Gritsenko, D., Wijermars, M., Kopotev, M. (eds.) Palgrave Handbook of Digital Russia Studies. Palgrave Macmillan, Basingstoke (2021)
15. Juva, M.: Valtiokirkosta kansankirkoksi: Suomen kirkon vastaus kahdeksankymmentäluvun haasteeseen. WSOY, Porvoo (1960)
16. Kokko, H.: Suomenkielisen julkisuuden nousu 1850-luvulla ja sen yhteiskunnallinen merkitys. Historiallinen Aikakauskirja **117**(1), 5–21 (2019)
17. La Mela, M., Tamper, M., Kettunen, K.: Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) DHN 2019 - Digital Humanities in the Nordic Countries. pp. 295–307. CEUR Workshop Proceedings, CEUR (2019), https://cst.dk/DHN2019/DHN2019.html
18. Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., Findlater, L.: The human touch: How non-expert users perceive, interpret, and fix topic models. International Journal of Human-Computer Studies **105**, 28–42 (Sep 2017). https://doi.org/10.1016/j.ijhcs.2017.03.007, https://linkinghub.elsevier.com/retrieve/pii/S1071581917300472
19. Li, Y., Nair, P., Wen, Z., Chafi, I., Okhmatovskaia, A., Powell, G., Shen, Y., Buckeridge, D.: Global surveillance of covid-19 by mining news media using a multi-source dynamic embedded topic model. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–14 (2020)
20. Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., Lahti, L.: Interdisciplinary collaboration in studying newspaper materiality. In: Krauwer , S., Fišer, D. (eds.) Twin Talks Workshop at DHN 2019. pp. 55–66. CEUR Workshop Proceedings, CEUR-WS.org, Germany (2019)
21. Marjanen, J., Pivovarova, L., Zosa, E., Kurunmäki, J.: Clustering ideological terms in historical newspaper data with diachronic word embeddings. In: 5th International Workshop on Computational History, HistoInformatics 2019. CEUR-WS (2019)
22. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., Tolonen, M.: A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. Journal of European Periodical Studies **4**(1), 54–77 (2019)
23. McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., Vatri, A.: A computational approach to lexical polysemy in Ancient Greek. Digital Scholarship in the Humanities **34**(4), 893–907 (2019)
24. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century american newspaper. Journal of the American Society for Information Science and Technology **57**(6), 753–767 (2006)
25. Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., Domínguez, L.M., Parker, J.: Spreading News in 1904: The Media Coverage of Nikolay Bobrikov's Shooting. Media History **26**(4), 391–407 (Oct 2020). https://doi.org/10.1080/13688804.2019.1652090, https://www.tandfonline.com/doi/full/10.1080/13688804.2019.1652090
26. Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q., McGillivray, B.: GASC: Genre-aware semantic change for Ancient Greek. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. pp. 56–66. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4707, https://www.aclweb.org/anthology/W19-4707

27. Pääkkönen, J., Ylikoski, P.: Humanistic interpretation and machine learning. Synthese (Sep 2020). https://doi.org/10.1007/s11229-020-02806-w, http://link.springer.com/10.1007/s11229-020-02806-w
28. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
29. Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., Ginter, F.: The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–15 (Sep 2020). https://doi.org/10.1080/01615440.2020.1803166, https://www.tandfonline.com/doi/full/10.1080/01615440.2020.1803166
30. Schofield, A., Mimno, D.: Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics **4**, 287–300 (2016)
31. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. pp. 63–70 (2014)
32. Sorvali, S.: "Pyydän nöyrimmästi sijaa seuraavalle" – Yleisönosaston synty, vakiintuminen ja merkitys autonomian ajan Suomen lehdistössä. Historiallinen Aikakauskirja **118**(3), 324–339 (2020)
33. Syrjämäki, S.: Sins of a historian: Perspectives on the problem of anachronism. Ph.D. thesis, Tampere University Press, Tampere (2011), oCLC: 816367378
34. Thompson, L., Mimno, D.: Authorless topic models: Biasing models away from known structure. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3903–3914. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1329
35. Tommila, P., Landgrén, L.F., Leino-Kaukiainen, P.: Suomen lehdistön historia 1. Sanomalehdistön vaiheet vuoteen 1905. Kustannuskiila, Kuopio (1988)
36. Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., Ginter, F.: Applying BLAST to text reuse detection in finnish newspapers and journals, 1771-1910. In: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. pp. 54–58 (2017)
37. Viola, L., Verheul, J.: Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. Digital Scholarship in the Humanities (2019)
38. Walker, D., Lund, W.B., Ringger, E.: Evaluating models of latent document semantics in the presence of ocr errors. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 240–250 (2010)
39. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 424–433 (2006)
40. Yang, T.I., Torget, A., Mihalcea, R.: Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 96–104 (2011)