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

D3.4: Final cross-lingual context and opinion analysis technology (T3.1)

Executive summary

This deliverable summarises the progress and outputs achieved on Task T3.1 of the EMBEDDIA project. Task T3.1 aims to develop cross-lingual tools for automatic analysis of the content and context of user-generated comments in news media, for use in Tasks T3.2 (comment analysis) and T3.3 (comment reporting). We describe our progress in developing new classifiers for the analysis of a range of relevant aspects, including author characteristics, opinion and sentiment (in each case summarising work reported at month M18 in the previous deliverable for this task, D3.2, and progress since then), and on two new areas: topic modelling, and detection of fake news and misinformation spreading. We show that our methods are suitable for cross-lingual transfer, and that the transfer methods developed in WP1 can be used to transfer our models to less-resourced EMBEDDIA languages, and that for some tasks this can be achieved with no measurable performance drop. Finally, we outline some ongoing work on stance detection that will lead to final results to be evaluated in D3.7.

Partner in charge: QMUL

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:	**D3.4**
Deliverable full title:	**Final cross-lingual context and opinion analysis technology**
Deliverable short title:	**Final cross-lingual comment analysis**
Document identifier:	**EMBEDDIA-D34-FinalCrosslingualCommentAnalysis-T31-submitted**
Lead partner short name:	**QMUL**
Report version:	**submitted**
Report submission date:	**30/06/2021**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Matthew Purver (QMUL), Ravi Shekhar (QMUL)**
Co-author(s):	**Boshko Koloski (JSI), Marko Robnik-Šikonja (UL)**
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
21/05/2021	v0.1	Matthew Purver (QMUL)	Initial draft.
28/05/2021	v1.0	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	First complete draft.
17/06/2021	v1.1	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	Submitted for internal review.
22/06/2021	v1.2	Jose G. Moreno (ULR)	Internal review.
24/06/2021	v1.3	Saturnino Luz (UEDIN)	Internal review.
25/06/2021	v1.4	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	Revision based on internal reviews.
27/06/2021	prefinal	Nada Lavrač (JSI)	Report quality checked.
29/06/2021	final	Matthew Purver (QMUL)	Report finalized.
30/06/2021	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DTD	Document-Topic Distribution
DTE	Document-Topic Embedding
ELMo	Embeddings from Language Models
ETM	Embedded Topic Model
IAC	Internet Argument Corpus
IRC	Internet Relay Chat
LASER	Language Agnostic SEntence Representations
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
mBERT	Multilingual BERT
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Neural Network
NYT	New York Times
POS	Part Of Speech
RNN	Recurrent Neural Network
UGC	User-Generated Content
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine

1 Introduction

The EMBEDDIA project aims to improve cross-lingual transfer of language resources and trained models using word embeddings and cross-lingual technologies, with a focus on nine languages: Croatian, English, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovene, and Swedish. Work package WP3 aims to apply EMBEDDIA's cross-lingual advances to help news media companies better serve their audience by understanding and analysing their reactions, and assuring the safety, fairness and integrity of their participation in public internet spaces. In Task T3.1, the focus is on automatic analysis of user-generated content (UGC) — primarily the comments readers post under news articles — and the context in which it appears.

The overall objective of workpackage WP3 is to apply EMBEDDIA's cross-lingual technologies to understand and analyse the reactions of multilingual news audiences. The specific objectives of WP3 are as follows:

- O3.1 Advance cross-lingual context and opinion analysis, via Task T3.1;
- O3.2 Develop cross-lingual comment filtering, via Task T3.2;
- O3.3 Develop techniques for report generation from multilingual comments, via Task T3.3.

The objective of this task, T3.1, is therefore to develop general cross-lingual methods for analysing the content and context of user-generated comments, for use in the filtering technologies developed in Task T3.2 and the summarisation methods to be developed in Task T3.3. To this end, we have developed a range of classifier models for short text classification, whose outputs can be used directly in Tasks T3.2 and T3.3, or re-trained for more specific versions of those tasks. We have investigated models using both conventional statistical models and deep neural networks (DNNs). The latter include a range of specific architectures built on context-dependent embeddings including BERT-based DNNs, suitable for integration with WP1's models for cross-lingual transfer.

The main contributions presented in this report (in the order of appearance) are as follows:

- Improved and more efficient classifiers for author profiling (including age and gender detection) via social network context, ranked 2nd in the PAN@CLEF author profiling shared task (Koloski et al., 2020a).
- Methods for incorporating multi-lingual topic models to improve the performance and interpretability of comment classifiers (Zosa et al., 2021).
- Classifiers to detect the spreading of fake news and misinformation, ranked 3rd in the PAN@CLEF shared task (Koloski et al., 2020b) and scoring within 1.5% of the top entry in the CONSTRAINT 2021 shared task (Koloski et al., 2021).
- Cross-lingual classifiers for sentiment detection, using WP1's cross-lingual models to transfer between EMBEDDIA languages with no measurable performance drop (Robnik-Šikonja et al., 2021).
- Multi-lingual classifiers for opinion detection in social media (Tabak & Purver, 2020),
- An emerging new multi-lingual dataset for stance detection in news comments, with initial cross-lingual classifiers based on WP1's models (Shekhar et al., 2021).

This report is split into 6 further sections. Section 2 summarises related work in analysing news comments and other user-generated content (UGC). In Section 3, we describe our work on analysis and fusion of various aspects of the context in which UGC text appears, including properties of the author, and topic of discussion. Section 4 describes our new work on detecting authors who spread fake news and misinformation in UGC, with models giving performance competitive with the state of the art in multiple languages. Section 5 describes our classifiers for detection of sentiment, opinion and stance, including experiments to show cross-lingual transfer potential (training on datasets in well-resourced languages and testing on others). Section 7 then summarises the main concrete outputs of this work,

and Section 6 summarises our conclusions and main findings, and outlines the connection to other ongoing EMBEDDIA tasks. The appendices include the papers on which the main content sections are based.

2 Background

In this section, we give an overview of the main analysis tasks providing the primary components for the applications developed in Tasks T3.2 (comment filtering) and T3.3 (comment summarisation and reporting).

2.1 Author analysis

Analysing characteristics of the authors of comments, or profiling them according to particular categories, can provide important information on which to base a summary or report of commenter behaviour (as will be developed in Task T3.3). Reports which give insight into the differences in opinion expressed by different age groups, or different genders, for example, can help news media publishers to understand their audience and how it varies across these segments. Profiling of this kind can also provide information to help improve further analysis via other natural language processing (NLP) tasks. Hovy & Søgaard (2015) show that many standard datasets, when used to train NLP tools, bias them towards the language of older people (not just in terms of vocabulary, but other aspects including grammatical structure), and give corresponding reductions in accuracy when applied to language from other age groups. Demographic information about authors can also help give better understanding of social media posts in a hate speech detection task (MacAvaney et al., 2019), and so may be key in achieving good performance in comment filtering for automated moderation (being developed in Task T3.2).

Author profiling has its roots in stylometric work and corpus analysis, e.g., the influential work of Koppel et al. (2002) in gender prediction showing that women have a more relational writing style (e.g., using more pronouns) and men a more informational style (e.g., using more determiners). Recent work has moved this into the computational NLP arena via shared tasks (e.g., Rangel et al., 2017, 2018) and corpora (e.g., Verhoeven et al., 2016). Much of this work is based on social media (e.g., Twitter) data, many recent examples are multilingual (with e.g., Verhoeven et al. (2016)'s TwiSty corpus covering six languages), and some tasks include cross-genre evaluation (e.g., Rangel et al., 2016; Dell'Orletta & Nissim, 2018); the methods and results achieved in such tasks therefore seem relevant to our task of UGC analysis, and to our multi-lingual setting and objective of cross-lingual transfer in EMBEDDIA. As far as we are aware, no public datasets supporting author analysis for specifically news comment UGC yet exist, so our work so far focuses on social media data. Social media and news comments share many features: both typically consist of UGC in short text form, and with a tree-like thread structure in which messages can respond to previous messages. As such, they tend to have similar linguistic properties, with abbreviated and informal language and many context-dependent phenomena. Many differences can be found too: social network platforms tend to have specific conventions (e.g. use of hashtags or automatic links) that might not appear in news comments; and news comments might be more topic-specific, given their relation to a specific news article. However, social media provides the closest type of data in the absence of suitable comments datasets, and the commonalities are substantial enough to enable good transfer learning from social media to news comments (see e.g., Pelicon et al., 2021, and the forthcoming Deliverable D3.6 for the task of offensive language detection).

Most approaches rely on vocabulary, typically using bag-of-words features and support vector machine (SVM) classifiers. The PAN 2017 gender-prediction competition winner used a SVM with very simple word 1-to-2-grams and character 3-to-5-grams (Basile et al., 2017); for age prediction, the PAN 2016 winners again used a SVM, this time with a broader range of features (word, character and POS n-grams, capitalization, punctuation, length, vocabulary richness, emoticons etc.) (Busger op Vollenbroek et al., 2016). Neural networks have also been applied; see e.g., (Miura et al., 2017) for experiments combining recurrent neural networks (RNNs) with convolutional neural networks (CNNs) together with

an attention mechanism. In our work so far, we have followed these standard approaches to produce systems with competitive accuracy on standard tasks.

2.2 Fake news and misinformation analysis

A new focus in this deliverable is the detection of UGC that attempts to spread fake news or misinformation. Most work on fake news text classification to date looks at social media data. The early proposed solutions to this problem used hand crafted features of the authors such as word and character feature distributions. Interactions between fake and real news spread on social media gave the problem of fake news detection a network-like nature (Shu et al., 2019). This network-based modelling discovered useful components of the fake news spreading mechanism and therefore approaches can overlap somewhat with the above-mentioned task of detection of bot accounts (Shao et al., 2018). Most current state-of-the-art approaches for text classification leverage large pre-trained models such as BERT (Devlin et al., 2019), and have promising results for detection of fake news (Jwa et al., 2019). Here, we investigate the combination of approaches: using network-like features as well as linguistic features, and investigating their fusion with features from large pre-trained models.

2.3 Sentiment and opinion analysis

Sentiment analysis and opinion mining have a long history in NLP and have become standard tasks for text processing (see Pang & Lee, 2008). However, the umbrella term *sentiment analysis* is often used to cover a range of more specific sub-tasks:

- *Subjectivity* analysis: determining whether a text contains subjective views or opinions or is purely objective/factual;
- *Sentiment* analysis: determining whether a subjective text expresses positive or negative sentiment;
- *Target-based* or *aspect-based* sentiment analysis: determining the positive/negative direction and/or strength of sentiment towards a particular target (usually an individual or organisation) or aspect of something discussed (e.g., the *plot* or *script* of a movie being reviewed, the *lens* or *price* of a camera);
- *Emotion* analysis: characterising the emotional content of a text, often categorising it along multiple dimensions according to primary emotions (e.g., happiness, sadness, anger, fear, disgust, surprise (Ekman, 1972));
- *Opinion* analysis: determining the author's stance (often: *affective* stance) or opinion on a particular subject.

The precise definitions and the desired level of analysis depend on the motivations and requirements of the research or application in question. In financial research, determining whether an article implies positive or negative sentiment towards a particular company's share price might be the overall objective. In news media UGC, our interest is likely to be directed towards determining the stance/opinion of users towards particular entities or topics. This will be a core component for other tasks: detecting opinions and their stance will be a fundamental component of the technology developed in Task T3.3 (report generation), and detecting negative emotions will be a component of the developments in Task T3.2 (comment filtering).

Again, the tasks above are generally approached as classification tasks, either binary or multi-class: see e.g., (Kalchbrenner et al., 2014; Müller et al., 2017; Li et al., 2018) for recent DNN approaches to sentiment analysis including target-based versions, (Purver & Battersby, 2012) for multi-class emotion detection in the Twitter domain using simpler linear SVMs, and (Celli, Stepanov, Poesio, & Riccardi, 2016) for opinion mining in online web comments (including news UGC).

Being a standard task, sentiment analysis has been applied to a wide range of datasets (see WP4 Deliverable D4.4 for discussion of sentiment analysis applied to news articles); but again, the most relevant work to our UGC domain is mostly on social media data. Several recent public shared tasks provide datasets annotated for sentiment and stance: for Twitter, Rosenthal et al. (2017) give an overview of recent years' sentiment tasks in the SemEval series and compile the datasets into one multilingual (English and Arabic) set, with annotation provided for simple sentiment polarity and for target. More recent tasks have started to address more specific sub-phenomena, with Mohammad et al. (2018)'s dataset for detecting *intensity* of multi-class emotion and sentiment, and Das et al. (2020) focussing on sentiment in code-mixed language, for example. For news comments, Celli et al. (2014) provide a small corpus annotated for sentiment (positive/negative polarity together with target topic) as well as emotion towards other comment posts (appreciation towards a message topic). Given their size, breadth and multilingual nature, we focus on standard social media data for now, and plan to apply the resulting methods and classifiers to news comment data in the implementation of Tasks T3.2 and T3.3.

2.4 Context analysis

One characteristic property of UGC in a news media context is that it consists of individual comments written by readers, but which are posted and read in an emerging context. Not only are the comments produced and consumed in the context of the news article to which they are attached, they appear in the context of other comments already present, and they then extend that context for the comments which may appear later. In this way, comments sections often have many of the properties of multi-party conversations: individual comments can refer to and build on other comments, and in turn be referred to and built on themselves. Success in many of our analytical tasks here will therefore depend on, or be improved by, the ability to model and incorporate contextual information from articles (and their multimodal content, including images and captions as well as text) and the ongoing conversation threads.

Work in multimodal text understanding was rare for many years, but has made good recent progress via the use of DNNs. In visual question-answering, for example, most successful methods use DNNs to fuse image processing with linguistic description (see e.g., Shekhar et al., 2019). In one of our specific tasks here, author profiling, multimodal tasks have been proposed, e.g., the multimodal gender classification task at PAN 2018 (Rangel et al., 2018) for gender prediction from Twitter texts combined with images. In this task, deep learning approaches prevailed with the overall winners using RNNs for texts and CNNs for images (Takahashi et al., 2018). In the news domain, Ramisa et al. (2018) show that CNNs can help fuse image and news text information in tagging and linking tasks; and Batra et al. (2018) combine CNNs and RNNs to generate captions for images in articles. We build on this work and investigate DNN methods for general context fusion.

Conversation thread modelling is also a key component: accuracy in tasks such as sentiment and opinion detection in news comments, and the comment filtering in Task T3.2, can be improved by the ability to automatically detect conversation structure and suitably model the ongoing context. For example, a comment C_2 in which the author agrees with a previous comment C_1 may be an example of hate speech if C_1 is an example of hate speech, but not otherwise. Comment C_2 may express a positive opinion or a negative opinion depending on the opinion expressed in C_1, and it may contribute to different topics depending on the content of C_1. Understanding agreement and disagreement relations has therefore proved to be important in previous work on summarisation of news comments (Barker & Gaizauskas, 2016), and on understanding opinions in online comments (Celli, Stepanov, Poesio, & Riccardi, 2016), both of which will become crucial in Task T3.3 (comment summarisation and reporting). Characterisation of this task varies, with most existing approaches examining sub-tasks such as agreement detection or antecedent detection, and seeing them as standard binary classification tasks (see e.g., Celli, Stepanov, & Riccardi, 2016, on news comment analysis). Tree- or graph-based variants can also be used, requiring different approaches to annotation and evaluation (see e.g., Zubiaga et al., 2016, when tracking rumours on Twitter).

Most work in this area is not directly applicable to our setting. Much work on thread structure is in

the domain of spoken two-person dialogue, which differs from our setting both in terms of conversation structure and language features. Some multi-party dialogue work is closer to our setting, and datasets suitably annotated for structure (including the presence of agreement and disagreement relations between utterances/speakers) include the ICSI and AMI corpora of multi-party meetings (Shriberg et al., 2004; Carletta et al., 2006). Corpora of written conversation also exist; these contain language phenomena which may be more similar to the UGC expected in our tasks, but come from a range of sources with different properties. The AAWD corpus contains messages from Wikipedia talk pages in multiple languages (Bender et al., 2011); the AACD chat corpus covers the same languages using text chat dialogue (Morgan et al., 2013); and the Internet Argument Corpus (IAC/ARGUE) corpus contains online forum political debates with over 11,000 conversation threads (Walker et al., 2012). Each is annotated with agreement and disagreement, with IAC also including labels for offensive language, sarcasm and attitude. Taking a slightly different perspective, Elsner & Charniak (2011) provide an annotated dataset of Internet Relay Chat (IRC) text chat dialogues in which discussion threads are interleaved between messages. While many properties may be different to news media comments, this shares the basic problem of distinguishing conversational relevance relations between messages. As might be expected given the varied nature of these datasets, modelling approaches vary widely.

Some data in the news domain exists, for example the German language Million Posts corpus (Schabus et al., 2017), but contains only very limited thread structure information. We are only aware of one dataset that contains more detailed annotations and is specifically from the online news UGC domain: the CoREa corpus (Celli et al., 2014) contains 27 news articles from the Italian online news site Corriere and the associated UGC comments, about 2,900 posts. Its small size makes it unsuitable for training and experiments here, so our approach so far focused on other data, with planned transfer to news data in later work.

2.5 Stance analysis and the role of context

One task in which context might play a particularly important role is that of opinion or stance analysis. Within the general field of sentiment analysis and opinion mining, most research takes the task to be one of text classification: determining the overall tone or stance of a text, with respect to some task-specific or domain-specific criteria (positive or negative opinion; author's emotional state; financial market outlook; etc.) Some tasks, however, are more focused, requiring stance towards some specific aspect or target – including recent SemEval tasks in which the stance of a text towards some given topic must be predicted (Rosenthal et al., 2017). However, although discussion between users about their stances towards given subjects is one of the primary uses of online forums and comment sections, there is little research so far that examines stances within an interactive context. Some recent examples of such research are (Zubiaga et al., 2018; Kumar & Carley, 2019) with a more detailed overview given in Küçük & Can (2020). On the other hand, while work in dialogue modelling often examines the interactive nature of agreement and disagreement between users, little of that examines how (dis)agreement comes together with the expression of stance.

Some recent datasets and tasks do take some interactive context into consideration. The Internet Argument Corpus 2.0 (Abbott et al., 2016), for example, uses dialogue structure to annotate multiple phenomena like topic, stance, and agreement between comments. Similarly, Allaway & McKeown (2020) propose a dataset in which comments from the New York Times (NYT) 'Room for Debate' are annotated with topics and stance. However, these approaches are designed to examine argument and thus focus on explicitly *controversial* topics like birth control, where strongly polarised stance is common. However, people frequently take and express stance on *non-controversial* topics. To investigate this, we use no specific topic-based data filtering, and consider stance towards more generic topics, using a richer annotation.

To capture a wider range of phenomena, we provide richer annotations including (dis-)agreement, target, stance direction and strength, and explanation of the annotator's decision. As there are few similar datasets for non-English languages (although some examples exist, see Bošnjak & Karan, 2019; Vamvas & Sennrich, 2020), we also provide annotated data in Croatian, to encourage work for lower

resourced languages or on cross-lingual approaches. We find additional motivation in recent findings on model explainability (summarized by Wiegreffe & Marasović (2021)), showing that explanations often require only shallow understanding of comments and no reasoning. This task is envisioned to require fine-grained complex analysis to generate explanations, gauging the potential of state-of-the-art NLP models.

3 Context analysis

One stream of work on this task has been on models with general architectures which extract useful information about the context of UGC text, including the social context (properties of the author) and discourse context (information contained outside the UGC text itself). By the time of the interim deliverable D3.2 we had developed a range of such models, and applied them to a range of tasks: detection of author gender and profile (Martinc & Pollak, 2019; Martinc et al., 2019b), distinguishing automated bots from human authors (Martinc et al., 2019a), understanding conversational context (Nasreen et al., 2019a,b) and integrating information from different modalities (Rohanian et al., 2019). Work since then has extended the approaches in author profiling (Section 3.1),and added a new direction by integrating topic modelling from WP4 (Section 3.2), resulting in a new set of tools providing useful information for more specific tasks within T3.1 (e.g. fake news detection, see Section 4), as well as downstream tasks such as filtering (Task T3.2) and summarisation (Task T3.3).

3.1 Author analysis

3.1.1 Work up to D3.2

As summarized in deliverable D3.2, work up to M18 resulted in a collection of accurate models for author analysis, tested in public shared tasks. For the CLIN 2019 shared task in gender classification, the task was monolingual (Dutch) but multi-genre; our system achieved 6th place in the within-genre setting, and 2nd place in the cross-genre setting (Martinc & Pollak, 2019). The PAN@CLEF 2019 shared task in Celebrity Profiling was again monolingual but provided a more complex multi-category task (gender, age and occupation); our system came 3rd in the final ranking (Martinc et al., 2019b). Also at PAN@CLEF 2019, the Bots and Gender Profiling task was multilingual (English and Spanish) and a very relevant task to WP3, bot detection; our system ranked 16th but used a similar model and performed within 4% absolute accuracy of the top-ranked entry (Martinc et al., 2019a).

This work therefore provided models giving state-of-the-art performance for core WP3 analysis tasks: gender, age and occupation detection, and bot detection. They were demonstrated to be applicable multi-lingually and to a range of tasks and genres. All were based on models that could be applied cross-lingually given aligned lexicons or embedding spaces. However, they were all based on the UGC text itself, and used no information from the context. Work since then has therefore focused on using the social context to see if this can provide similar or supplementary information.

3.1.2 Work since D3.2

In this work we approach the same task as before, of inferring author information (gender, age and occupation), but using information only from the social context: UGC published by users connected to the author in question (but not including the text written by the author themselves). The dataset and evaluation were provided by the PAN@CLEF 2020 Celebrity Profiling shared task.[1]

Features for classification are extracted from the user's social network: we select 20 tweets from each of 10 authors connected to the target user. A range of character-based and word-based ngram features

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

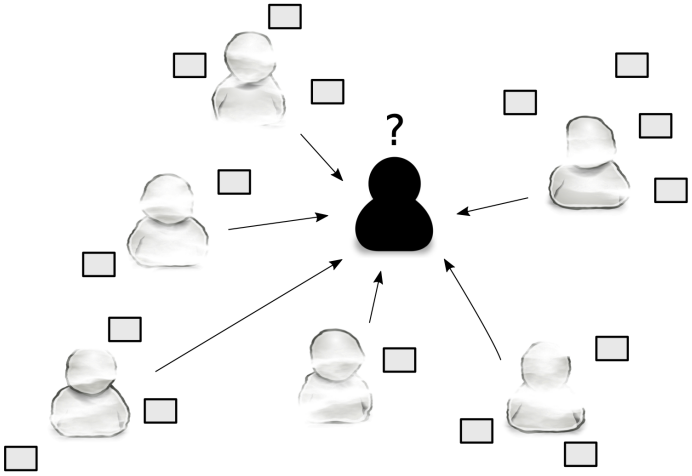


Figure 1: Schematic overview of the task and approach: tweets written by the target user's social network, rather than by the target user themselves (Koloski et al., 2020a).

are then extracted, and a SVD-based dimensionality reduction is performed (similar to that used in LSA (Landauer et al., 1998)). We tested a range of classifiers and feature selection methods, deciding on logistic regression for the age and occupation classes, and polynomial-kernel SVMs for gender. The best accuracies achieved on the test set were 40.7% for age, 61.6% for gender and 59.7% for occupation, and the system ranked 2nd overall in the combined ranking. Table 1 shows results in the different subtasks for a range of competitive classifiers and feature sets.

Table 1: Final evaluation on training dataset (Koloski et al., 2020a). AC, FC and R are different ways of representing the *age* variable: R is a continous regression model; FC is a multi-class classification into 60 one-year classes; AC is a multi-class classification into 8 age intervals. Bold shows the best result for each task.

name	#features	#dimensions	f1 age	f1 gender	f1 occupation	crank
model-AC-2	20000	512	**0.358**	**0.665**	0.656	**0.516**
model-AC-1	20000	512	0.346	0.663	**0.669**	0.509
model-FC-2	10000	512	0.313	0.639	0.632	0.473
model-FC-1	10000	512	0.291	0.605	0.648	0.452
model-R	10000	512	0.298	0.612	0.613	0.453
baseline-ngrams	#	#	0.362	0.584	0.521	0.469

The results show that this approach gives accuracies which are competitive with, although not as good as, a system which uses the text from the target user's tweets themselves (which would achieve 50.0% for age, 75.3% for gender, 70.0% for occupation). This suggests that this use of social context can give useful information which might be leveraged in a fuller system, and which is independent of language.

This work is described in full in (Koloski et al., 2020a), attached here as Appendix A.

3.2 Topic analysis

This stream of work is new since deliverable D3.2 (M18). Up to that point, extensive work on topic modelling had been carried out in WP4 (applied to analysis of news articles, see e.g. Zosa & Granroth-Wilding, 2019; Marjanen et al., 2021), but had not yet been applied to UGC (news comments). Since

then, a new collaborative strand of work has brought WP3 and WP4 together to apply and adapt the methods of WP4 to develop models of topic suitable for comment analysis.

We use the Embedded Topic Model (ETM, Dieng et al., 2020) as our topic model since it has been shown to outperform regular LDA (Blei et al., 2003) and other neural topic modelling methods such as NVDM (Miao et al., 2016). We also want to take advantage of ETM's ability to incorporate the information encoded in pretrained word embeddings trained on vast amounts of data to produce more coherent topics. In the ETM, topics are embedded in the same space as the word embeddings, and are learned during topic inference, while the word embeddings can be either learned or pretrained; we use pretrained embeddings.

We train this model on one of our EMBEDDIA media partner news comment datasets, the 24sata comment dataset (Shekhar et al., 2020). This contains c.21M comments on 476K articles from the years 2007-2019, written in Croatian. For a 100-topic model on the entire test data, the top topics of non-blocked comments cover coherent topics over a diverse range of subjects from politics to football to scientific research. Table 2 shows some of these topics (labels are manually assigned by native speaker) with English translations.

Table 2: Selected topics with English translations. The first two topics are prevalent in non-blocked comments, the next two are prevalent in blocked comments, and the last is prevalent in both classes.

Croatian football	dinamo, hajduk, zagreb, zagrebu, placu, europi, zagreba *(dynamo, hajduk, zagreb, zagreb, market, europe, zagreb)*
State and government	država, države, državi, vlasti, državu, vlade, vlada *(state, states, state, authorities, state, governments, government)*
Moderately offensive	gluposti, sramota, sram, glup, jadni, jadan, jadno, budale *(nonsense, shame, disgrace, stupid, miserable, miserable, miserable, fools)*
Death and illness	žena, žene, ljudi, osoba, osobe, ženu, smrt, čovjeka *(woman, women, people, person, persons, woman, death, human)*
Civil war	srbi, hrvata, tito, srba, srbije, srbiji, srbima, srbija *(serbs, croats, tito, serbs, serbia, serbia, serbs, serbia)*

More interestingly, from the point of view of downstream tasks such as the comment filtering task of T3.3, different topics show different associations with the likelihood that a human moderator would block a comment; and these associations vary across different sections of the newspaper. Figure 2 shows an example for two 24sata news sections (*Lifestyle* and *Politics*), across blocked and non-blocked comments (as determined by 24sata's moderators). Different topics can be seen to appear in different areas; some associations may be unintuitive, e.g. the association of the "Football cards" topic with blocking in the *Lifestyle* section, but turn out to be meaningful: commenters often discuss moderator's blocking decisions as "yellow cards" or "red cards", and this discussion is associated with further blocking decisions. The same association does not hold, of course, in the *Sports* section.

This leads to a potential for use in improving classifiers for automated comment moderation: topic distribution information should provide information useful for a blocking decision, beyond the immediate words of the comment. Given a trained ETM, we can infer the **document-topic distribution (DTD)** of an unseen document. In addition, we can also compute a **document-topic embedding (DTE)** as the weighted sum of the embeddings of the topics in a document, where the weight corresponds to the probability of the topic in that document. We then propose two alternative models to fuse this information with the comment text, similar to the fusion architecture introduced in our earlier work on context analysis (see Rohanian et al., 2019, and the earlier Deliverable D3.2) - see Figure 3. The use and evaluation of this approach in automated comment filtering (Task T3.2) will be presented in Deliverable D3.6.

This work is described in full in (Zosa et al., 2021), attached here as Appendix B.

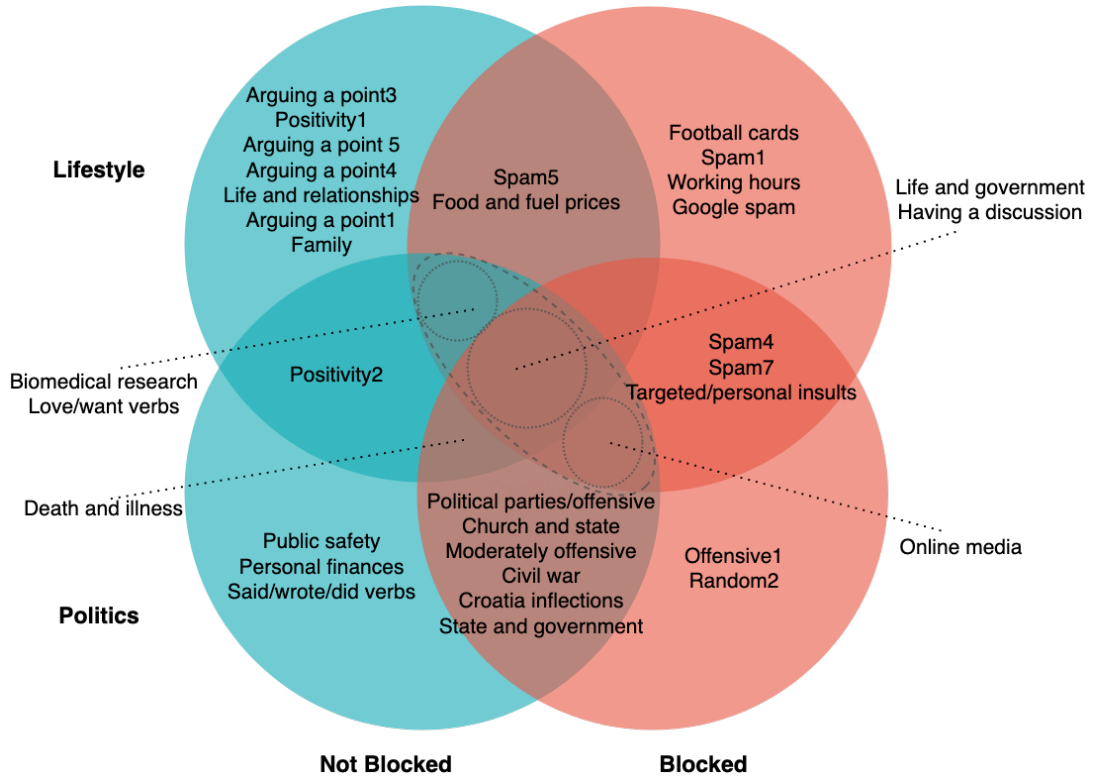


Figure 2: Top topics of the blocked and non-blocked comments for the entire test set (Zosa et al., 2021).

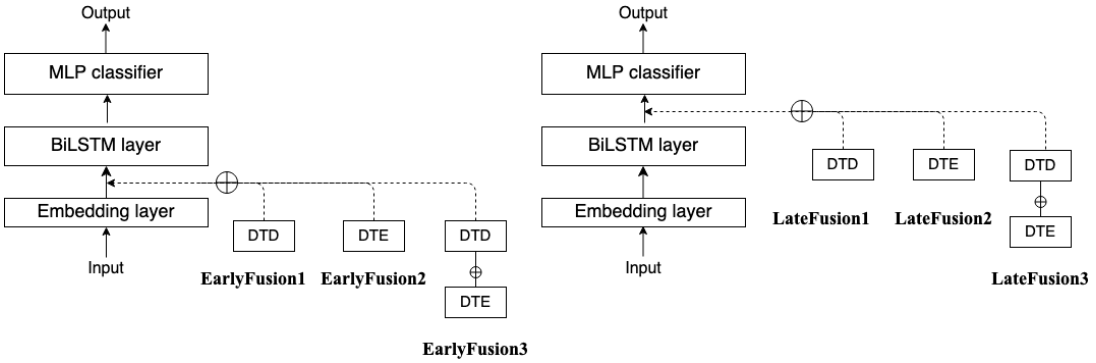


Figure 3: Network structure for topic/comment fusion (Zosa et al., 2021).

4 Fake news and misinformation analysis

Another entirely new stream of work that has become a focus since deliverable D3.2 (M18) is the development of models for detection of fake news and misinformation in comments. The task of misinformation detection has risen to prominence in recent years, and seems likely to become a subject of interest for news media companies. Already, one of the reasons for blocking of comments by moderators is the presence of misinformation (see Shekhar et al., 2020, and Deliverable D3.3); although our previous work therefore addressed it implicitly, as part of the comment filtering task, it now seems timely to investigate models which can specifically detect and analyse this important phenomenon. This was noted in reviewers' comments received at the mid-term EMBEDDIA review that:

> *[. . .] fake news is a very important topic in the project domain, and should be studied in depth [. . .] [Task T3.3] addresses offensive language, but it would be interesting to study the impact and advances*

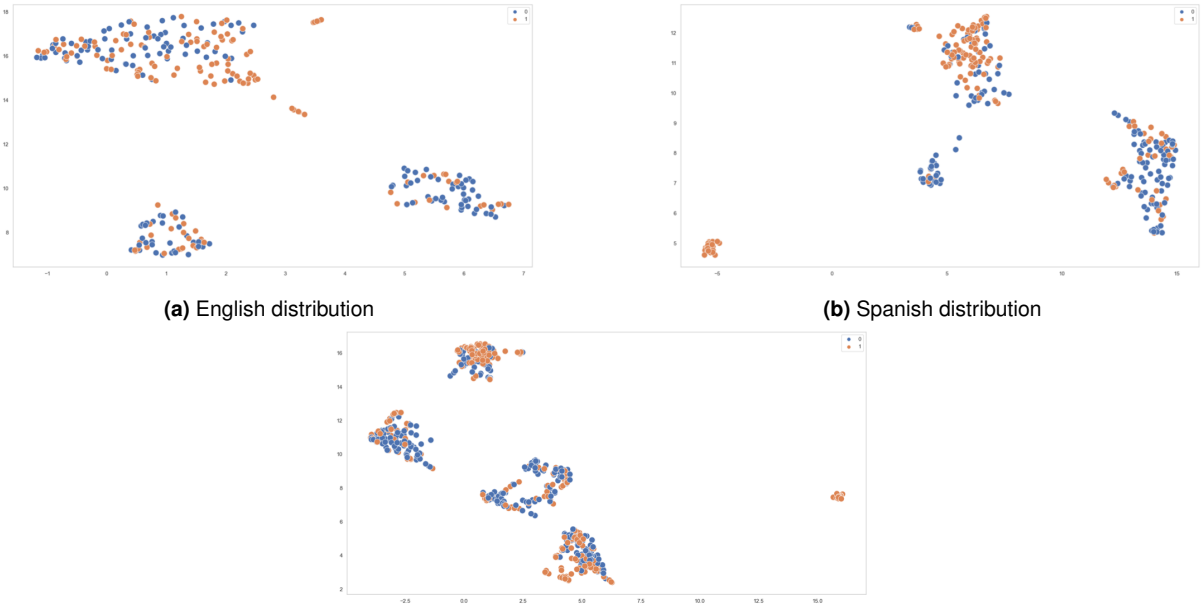
in fake news (one of the main problems in today's media).

We hope that this new focus addresses this issue.

4.1 Detecting fake news spreaders

Our first work in this direction took the general author-profiling approach, as described in Deliverable D3.2 and Section 3.1 above, and applied it to the task of detecting fake news spreaders. We developed a system for the PAN@CLEF 2020 shared task on Profiling Fake News Spreaders on Twitter (Rangel et al., 2020): given a timeline of tweets from a particular author, the goal is to decide if the author is a spreader of fake news or not. The task was multilingual, run in English and Spanish, with a dataset of 100 tweets from 300 authors in each language (150 fake news spreaders, 150 not).

We followed a similar approach to the one we used for author profiling in (Martinc et al., 2019b) (see D3.2), deriving a large set of character- and word-based ngrams from the tweet texts and testing a range of standard classifiers (logistic regression, SVMs, random forests). We used dimensionality reduction to give a low-dimensional representation of the feature space, in a similar way to our author-profiling approach in Section 3.1 above. We also tested two approaches to the multilingual nature of the task: a monolingual approach in which the data for each language was treated separately, training separate English and Spanish classifiers; and a multilingual approach in which the data from both languages was fused. As Figure 4 shows, a projection of the data into the dimensionally-reduced latent feature space suggests that the data from the two languages can be usefully clustered together, suggesting that a multilingual approach can give benefits.



(a) English distribution

(b) Spanish distribution

(c) Merged distribution

Figure 4: UMAP visualization of the latent spaces used to train the final models. The orange colour corresponds to spreader and the blue to non-spreader. The plots indicate the number of clusters is maintained in the latent space. (Koloski et al., 2020b).

The performance on the training set was good (see Table 3), and suggested that while the multilingual models gave competitive performance, the monolingual models might have an advantage in some cases. However, the multilingual approach proved more robust: performance on the test set (shown in Table 4) shows that it equalled or outperformed the monolingual models. Our approach gave very competitive results, ranking 2nd overall in the PAN shared task.

Table 3: Final results on training dataset for a range of feature types and classifiers (Koloski et al., 2020b). Bold shows the best results for each language.

name	type	#features	#dimensions	model	EN ACC	ES ACC
tfidf_large	multi	5000	768		0.9633	**0.9867**
tfidf_tweet_tokenizer	multi	5000	768	LR	0.9633	0.9533
tfidf_small	mono	5000	512	SVM,SVM	**0.9700**	0.4900
tfidf_cv	mono	10000	768	SVM,SVM	0.9100	0.9367
tfidf_no_hash	multi	10000	768	LR	0.9300	0.9067
doc2vec_baseline	mono	100	#	RF,SVM	0.6428	0.6971
tfidf_tpot_baseline	mono	30000	#	LR,SVM	0.7500	0.7400
tfidf_baseline	mono	10000	#	LR,LR	0.5567	0.7033

Table 4: Evaluation on test dataset for final models (Koloski et al., 2020b). Bold shows the best results for each language.

name	type	#features	#dimensions	model	EN ACC	ES ACC
tfidf_large	multi	5000	768	LR	**0.7150**	**0.7950**
tfidf_cv	mono	10000	768	SVM,SVM	0.7000	**0.7950**

This work is described in full in (Koloski et al., 2020b), attached here as Appendix C.

4.2 Neural models for fake news

The approach described in the previous section, while providing good levels of accuracy in monolingual and in multilingual settings, uses classifiers that although deployable cross-lingually given a suitably cross-lingually aligned lexicon, are not directly compatible with the cross-lingual methods developed in WP1. WP1's methods are based on deep neural networks (DNNs), particularly large pre-trained language models such as BERT (Devlin et al., 2019) - see e.g. (Ulčar & Robnik-Šikonja, 2020). In our next steps, therefore, we examined the use of these large pre-trained language models in the task of fake news spreader detection, with a view to combining them with WP1's models in future.

Our approach in this stage was to try to combine the advantages of hand-crafted and lexical features (seen to perform well in our previous work, see Section 4.1) with the general, robust contextual embedding representations provided by large pre-trained language models. We therefore used a *stacking* architecture in which lexical features (dimensionally reduced via LSA) could be combined with a range of contextual embeddings - see Figure 5. As lexical features, we used word- and character-based ngrams (as before), together with POS information and hand-crafted length/number features (e.g. maximum/average/minimum word length in a tweet, standard deviation of word length, number of words beginning with upper/lower case). For contextual embeddings, we used a range of models including DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), XLM (Lample & Conneau, 2019) as well as tax2vec, a model which incorporates background knowledge from the WordNet taxonomy (Škrlj et al., 2021). Our stacking approach then combined individual classifier outputs via either stochastic gradient descent (SGD) (*linear* stacking) or a 5-layer neural network (*neural* stacking).

For this approach, we tested and evaluated our models on the CONSTRAINT 2021 workshop shared task on *COVID19 Fake News Detection in English*. The dataset consists of social media posts in English collected from Facebook, Twitter and Instagram, and the task is to determine for a given post if it was real or fake in relation to COVID-19. Performance was good, with final F1 score on the unseen test set of 97.2%. The system's ranking was only 50th place amongst 168 submissions, but its performance was within 1.5% of the best performing solution.

This work is described in full in (Koloski et al., 2021), attached here as Appendix D.

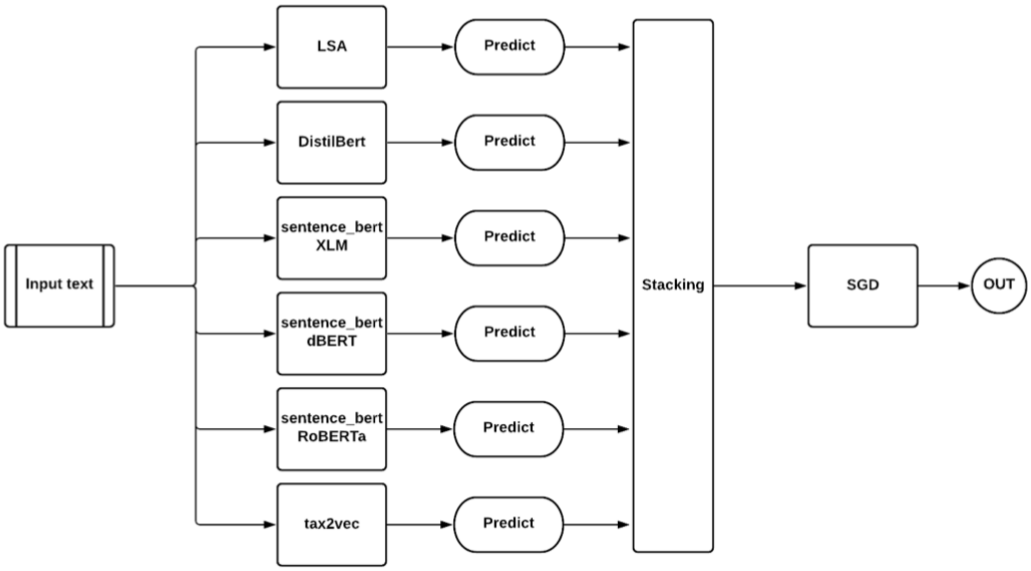


Figure 5: Linear stacking architecture used to combine the base classifier models. The neural stacking approach uses the same structure, and substitutes a 5-layer deep neural network for the SGD component (Koloski et al., 2021).

Table 5: Final F1-score results for our shared task submissions (Koloski et al., 2021). Bold shows the best result.

submission name	model	F1-score
btb_e8_4	neural stacking	**0.9720**
btb_e8_3	LSA	0.9416
btb_e8_1	tax2vec	0.9219
btb_e8_2	linear stacking	0.8464
btb_e8_5	distilbert	0.5059

5 Sentiment and opinion analysis

The final thread of UGC content analysis work has been a continuation of work before M18, focused on detection of sentiment and opinion, key building blocks likely to be required for a useful summary of news comments in Task T3.3 (see e.g., Riccardi et al., 2016; Barker & Gaizauskas, 2016). By the time of the interim deliverable D3.2 we had developed models for opinion and sentiment detection in English (Concannon & Purver, in preparation), and investigated initial cross-lingual transfer for sentiment detection across a range of languages including many EMBEDDIA languages (Robnik-Šikonja et al., 2020). This work showed that practical models could be built, but in terms of opinion and sentiment phenomena was limited to explicit opinions expressed in single messages, and limited to social media data (Twitter). The cross-lingual work showed that cross-lingual transfer could be practical in some cases, but with a drop in accuracy from the monolingual case; however, it only examined the use of large, multilingual pre-trained models available in standard libraries (specifically, BERT (Devlin et al., 2019) and LASER (Artetxe & Schwenk, 2019)).

In our recent work, therefore, we have extended this thread in three main ways: to look at news comments specifically, rather than social media data; to examine a wider range of phenomena, including implicit sentiment expressed through reactions, and opinions expressed in conversational thread contexts rather than individual isolated texts; and to integrate WP1's progress in cross-lingual modelling to improve cross-lingual transfer. The following sections describe these in reverse order.

5.1 Sentiment detection and cross-lingual transfer

5.1.1 Work up to D3.2

In our work up to D3.2 (M18), we examined the use of multi-lingual pre-trained models and transfer learning to enable cross-lingual transfer of sentiment classifiers. We used UGC text from Twitter, training classifiers for a three-way positive/neutral/negative sentiment task, experimenting with 13 languages. We used pre-trained multilingual sentence encoders (BERT and LASER) in standard classification setups (passing LASER outputs through a densely-connected NN layer, and fine-tuning the last layer of BERT's Transformer stack). Using LASER, we compared the zero-shot cross-lingual transfer setting (training on a source language, testing on a different target language) with the ideal monolingual case (training and testing on the target language), and found an average performance drop of 9% absolute F1 score for languages in the same family, and 14% for languages in different families. We also investigated whether augmenting target-language training data with other languages could improve performance, but found that on average it made performance worse by 4% absolute F1 score.

5.1.2 Work since D3.2

Table 6: Classifier performance as macro-averaged F_1 score and classification accuracy (CA) for the three-way sentiment task with zero-shot cross-lingual transfer (training on *source* language data and testing on *target* language data), for all combinations of languages on which CSE BERT was trained (Croatian, Slovene, and English) (Robnik-Šikonja et al., 2020). Bold shows the best performing cross-lingual model for each source/target language combination.

		LASER		mBERT		CSE BERT		Both target	
Source	Target	F_1	CA	F_1	CA	F_1	CA	F_1	CA
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	English	**0.63**	0.63	**0.63**	**0.66**	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.5	0.53	**0.59**	**0.57**	0.60	0.60
English	Croatian	0.62	**0.67**	0.67	0.63	**0.73**	**0.67**	0.73	0.68
Slovene	English	0.63	0.64	**0.65**	**0.67**	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	**0.60**	**0.58**	0.60	0.60
Croatian Slovene	English	0.62	0.61	**0.65**	**0.67**	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	**0.68**	**0.70**	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

To investigate the possible improvements to be gained from different cross-lingual embedding models, and particularly by integrating WP1's results on cross-lingual modelling, we extended our experiments to examine not only LASER but BERT in two forms: the standard multilingual BERT, trained on 104 languages (Devlin et al., 2019, hereafter mBERT), and the EMBEDDIA WP1 CroSloEngual BERT, trained on Croatian, Slovene and English (Ulčar & Robnik-Šikonja, 2020, hereafter CSE BERT).

Again, we investigated the potential for the challenging case of zero-shot transfer: comparing the performance achieved by training purely on a single source language and testing on the target language, with that which would be achieved if we were able to train and test on target language data. These experiments showed that on average over many languages, the transfer performance was similar to the previous experiments, with an average performance drop for same-family transfer of 5% F1 score for LASER, 6% for mBERT and 8% for CSE BERT. However, when examining the EMBEDDIA languages of interest on which CSE BERT was trained, we see a very different picture: while LASER and mBERT still show significant performance drops of around 4% F1 score, CSE BERT shows no performance drop - see Table 6. In other words, when used for cross-lingual transfer between the languages for which it was designed, CSE BERT gives almost perfect zero-shot transfer: near-identical performance when training only on a different source language and with no target language training data at all.

We therefore see this approach as highly practical for the production of sentiment classifiers in languages or domains for which no annotated data is available, by using WP1's techniques for pre-training suitable language models on unannotated data, and training a cross-lingual classifier on an annotated dataset in a different source language. The classifiers and code have been made public, and contributed for use in Task T3.3's summarisation and reporting models; and the cross-lingual technique — and specifically the use of CSE BERT for our EMBEDDIA languages — is now used in Task T3.2's comment filtering models, to be described in Deliverable D3.6.

This work is described in full in (Robnik-Šikonja et al., 2021), attached here as Appendix E.

5.2 Opinion detection in social media

5.2.1 Work up to D3.2

Our previous work in this direction focused on the detection of UGC texts that contain explicit opinions, via a dataset of social media posts in a specific domain (the UK health service). We achieved reasonable performance using a BERT classifier model, with over 70% macro-averaged F_1 score on a highly unbalanced task (texts containing opinions made up only 1-2% of data) (Concannon & Purver, in preparation). However, this work was monolingual (English), domain-specific (healthcare tweets) and looked only at explicit opinion expression (tweets labelled as opinion-containing under a narrow annotation definition).

5.2.2 Work since D3.2

In work since M18, then, we have expanded the scope to confirm the practicality of more general opinion mining: detecting implicit opinion by observing reactions in general, open-domain text, and comparing performance across different languages, allowing us to examine differences in reaction in different countries.

We built a classifier to detect negative reactions in general text by using the distantly supervised method of Coppersmith et al. (2014): rather than using an explicitly annotated dataset, we collected tweet timelines from the Twitter API for a set of users who had recently declared themselves diagnosed for depression. These were taken to be representative of the language of depression (i.e. emotionally more negative), while a random sample of other users was taken as a control set; these were then used as the "positive" and "negative" labelled sets to train a standard DNN classifier (here, a BiLSTM). This achieved reasonable accuracy for such a challenging task on a noisily-labelled dataset: 0.69 macro-averaged F1 score. This process was repeated across multiple languages (English, Italian, Spanish, French, German) - the lack of requirement for manually annotated data makes it cross-lingually applicable.

This classifier was then used to monitor a set of general, open-domain Twitter streams over the first half of 2020, by randomly sampling from the general stream filtered by geolocation coordinates located in the countries of interest. This method allows us to examine whether we can track relative changes in negative reactions in the general population on Twitter (absolute levels are not interpretable). As Figure 6 shows, the outputs show significant changes in negative emotion rates over time, and these can be correlated with known events in the real world. Figure 6(a) for the UK shows a strong upturn in depression-related language on Christmas Day, a known effect. Comparing Figures 6(a-c), we can then see that reactions to different countries' COVID-19 lockdowns: for example, UK attitudes seemed to become increasingly negative as other countries locked down but the UK did not, easing once that UK policy changed; while attitudes in Italy became increasingly negative after their own lockdown, easing once other countries joined them.

This work therefore shows that classifiers to measure implicit opinions in general open domains can be developed with minimal annotated data, with this technique applied across languages and used to

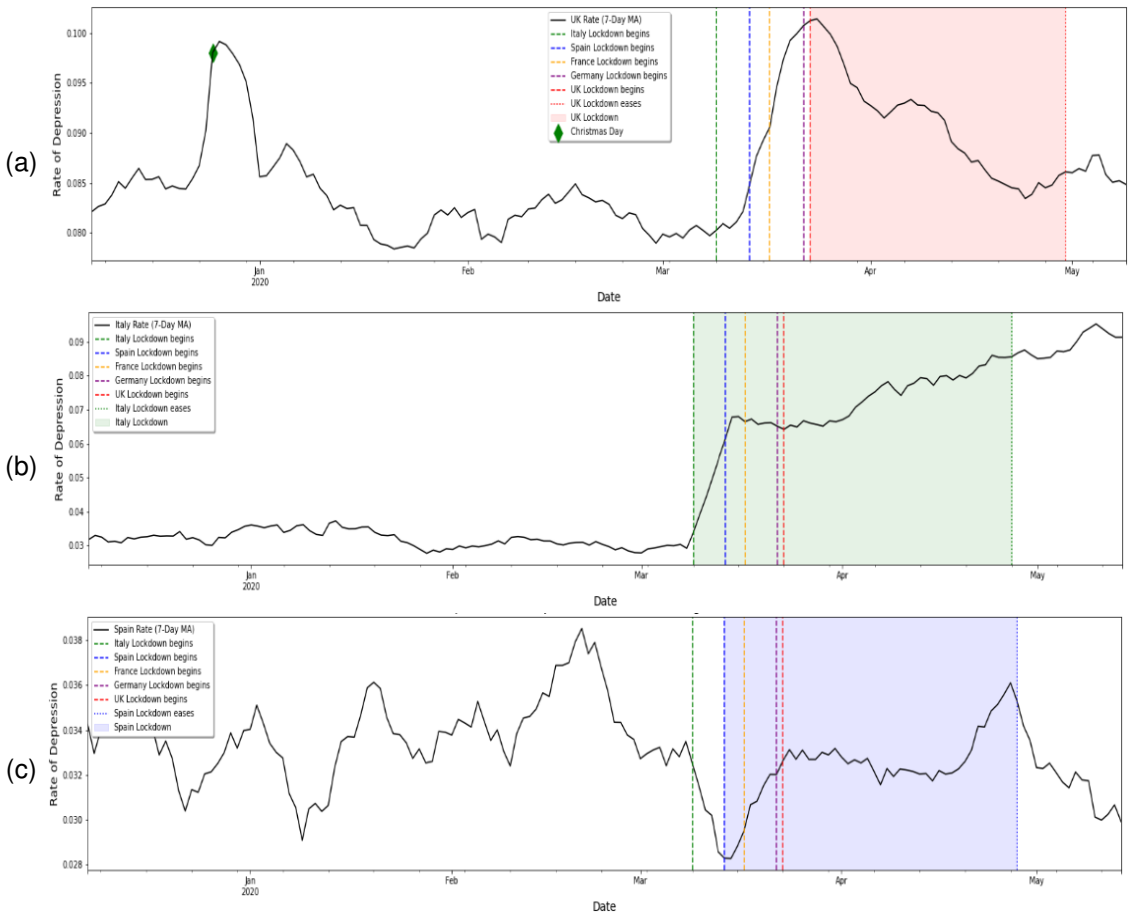


Figure 6: Variation of negative reactions, as measured by rate of language associated with depression, over time during various stages of national COVID-19 lockdown in (a) the UK, (b) Italy, (c) Spain (Tabak & Purver, 2020).

understand reactions to world events. This provides a way to develop general classifiers for use in UGC analysis and summarization, as required in in Task T3.3.

This work is described in full in (Tabak & Purver, 2020), attached here as Appendix F.

5.3 Stance detection in news comment threads

5.3.1 Work up to D3.2

When analysing the discussions in news comments, we need to examine not only sentiment (general positive/negative expression) and opinion on topics, but *stance* towards the opinions expressed by others: agreement/support or disagreement/opposition to other comments and/or the news articles themselves. Up to M18, our work in this direction concentrated first on thread reconstruction (detecting which previous comment in a thread is being responded to by some later comment), and next on stance classification. In the thread reconstruction task, we developed a cross-lingual classifier with reasonable accuracy by using the LASER encoder (Artetxe & Schwenk, 2019) with a simple NN classification layer, tested on news comments in German (via the One Million Posts corpus (Schabus et al., 2017)) and in Croatian (via the EMBEDDIA 24sata corpus (Shekhar et al., 2020)). For the stance detection task, we used a similar LASER-approach applied to antecedent-response comment pairs, achieving F1 score

of 0.80 on news comments in Italian (via the CoREa corpus (Celli et al., 2014)) and English (via the YNACC corpus (Napoles et al., 2017)). These results are very encouraging, but were achieved on small corpora, and not on the EMBEDDIA languages, due to the lack of data annotated for stance.

5.3.2 Work since D3.2

Work since D3.2 has therefore focused on extending this stance detection work, creating new datasets in EMBEDDIA languages to support that, and testing cross-lingual transfer.

Definitions As Barker & Gaizauskas (2016) point out, effective analysis of user comments requires us to understand user intent in a thread, and in particular whether a user's comment is intended to agree, disagree, or express neutral stance towards another user's comment or view. Understanding these different viewpoints could help in effectively summarising the overall discussion of the article. This could also assist in analysing opinion and public sentiment on a particular topic. A second part of our thread analysis problem is therefore to enrich the response-antecedent relations inferred with stance information: detecting whether the response is intended to agree or disagree with the antecedent. To capture these phenomena, we provide richer annotations including (dis-)agreement, target, stance direction and strength, and explanation of the annotator's decision; see Table 7 for examples.

Table 7: A sample example to illustrate the data annotation schema, taken from the New York Times (NYT) dataset. Note that in the NYT data, commenters usually mention the username of the person being responded to; this is not the case in the 24sata data. Note that the target of the opinions expressed is not directly mentioned in the comments.

#	Antecedent	Text	Target	Stance	Agreement
1	N/A	No labels : While running may be good for bone density, its certainly bad for your joints and, as you age, for your spine. So I suppose you have to pick your poison or, as many have observed here, do your sport in moderation.	running is good for you	-1	N/A
2	1	George Carlson : @No labels Unless you have some sort of skeletal defect or have an old injury from an accident or sport like football, studies show that running is not bad for your joints.	running is good for you	+1	-1
3	1	mrsg : @No labels I had a hip replacement two years ago and was told by the surgeon I could do pretty much anything I wanted except running and basketball. The repeated pounding wears out the joint cartilage, or in my case it would wear out the new polyethylene lining that now faces the titanium ball and socket. Repeated stress does wear out the joints.	running is good for you	-1	+1
4	1	Kim : I seem to be prone to get plantar fasciitis and now I've decided running is just not a form of exercise I can do anymore.	running is good for you	0	+1
5	4	Ron A : @Kim A foot doctor I once talked to told me she had PF but she wouldn't let it stop her. She was in the middle of running a marathon in every state in the US.	running is good for you	+1	-1

Tasks Our approach to the problem, and therefore the annotation process required to support it, sees it as four distinct subtasks:

Task A: (Dis-)Agreement Classification The simplest version of the task is to classify a given comment as agreeing or disagreeing with its antecedent comment. This can be framed as a classification task over pairs of comments: for any pair, predict the correct label from a three-way choice (agree, disagree, none/mixed).

Task B: Target Identification A more challenging task is then to predict for each comment a list of targets for which stance is being expressed (the *stance focus*, see Kiesling et al., 2018). Targets are often not mentioned in every comment, but instead must be inferred from the context: in most cases, we expect approaches that choose a key word or phrase from a comment somewhere in the

thread will be able to do well, but in some cases only more adventurous approaches that generate candidate phrases or choose them from the associated news article will be able to succeed.

Task C: Target-based Stance Identification Task B will then be followed by a task of identifying the stance direction and strength as regards each target, on a 5 point scale from strongly negative to strongly positive.

Task D: Explanation Generation The final task is generative: for each classification decision in Task A, systems must produce a short text explanation of their decision. This text is expected to include the key words/phrases in the comments that make the viewpoints and (dis)agreements clear, but may rephrase or reformulate them freely.

Dataset We use comments from two newspapers, the New York Times (NYT) in English and 24sata in Croatian (EMBEDDIA data). We collect the (dis)agreement between each comment and its antecedent and a free-text explanation of this decision. We annotated 50 articles from both datasets. Articles were selected such that it covers diverse range to the news section. For annotation, we recruited native speakers for each language.

Table 8: Data distribution for the New York Times (English) and 24sata (Croatian) datasets.

Source	Train		Val		Test	
	#articles	#comment	#articles	#comment	#articles	#comment
NYT (EN)	30	396	10	175	10	162
24Sata (HR)	30	720	10	321	10	279
Total	60	1116	726	496	730	441

Model In this deliverable, we only provide a model for Task A. We trained models on the English data and tested them on both English and Croatian data. For training, we used models based on the EMBEDDIA BERT (cseBERT, Ulčar & Robnik-Šikonja, 2020). For English data, the model achieved an F1 score of 0.72, while for the English-to-Croatian cross-lingual transfer model, performance is only 0.55. This shows that a model trained and tested on the same language achieves encouraging performance, but suggests that cross-lingual transfer is challenging. However, this may not just be due to linguistic differences, but could also be attributed to the nature of the comments in different newspapers: NYT comments are longer and highly informative, while 24sata comments are shorter and less formal. In the future, we plan to include these variations in the model; extensions to the model will be reported in Deliverable D3.7.

This work is described in full in (Shekhar et al., 2021), submitted as a proposal for a task at SemEval-2022, attached here as Appendix G.

6 Conclusions and further work

The objective of this task T3.1 was to develop effective analysis methods for use in cross-lingual UGC tasks, particularly for use in Task T3.2 (news comment filtering) and Task T3.3 (news comment summarisation). As Sections 3, 4 and 5 show, we have succeeded in developing classifiers for a range of suitable analysis tasks: author profiling, topic modelling, fake news detection, sentiment, opinion and stance analysis. New directions since the interim deliverable D3.2 include the successful integration of topic modelling and fake news detection. Our models have all now been tested on EMBEDDIA project languages, and all use methods that can be combined with WP1's results on cross-lingual embeddings to produce cross-lingual versions for transfer to new less-resourced languages. As shown in Section 5, for some analysis tasks this cross-lingual transfer can be achieved very well, with no measurable drop in accuracy even in the zero-shot case (Section 5.1), while in other cases the differences in data or genre can make it much more challenging (Section 5.3).

As Task T3.1 finishes, we are now transferring the models and methods developed for use in comment filtering (Task T3.2) and summarization (Task T3.3), and will continue to extend our stance analysis dataset and models for final evaluation in Task T3.4. We are also integrating selected models into WP6's Media Assistant implementation, and intend to work with media partners towards end-user testing in commercial media environments.

Multimodal analysis One specific area in which we hope to extend our work on T3.1 in a new direction is that of multimodal fusion. As summarized in the earlier interim deliverable D3.2, some of our work up to M18 focused on developing a general multimodal model capable of learning the relations between information sources that may be in different modalities and have complex temporal or sequential relations — with the intention of using this to model the relation between the user comments and various aspects of news articles, including the associated text, images and video material etc.). This led to a novel gated neural network architecture, structured to allow fusion of information from sources in different modalities, with different structures (Rohanian et al., 2019). Since M18 this has been refined and applied to different tasks, showing significant gains in accuracy as information from multiple modalities were combined (Rohanian et al., 2020, attached here as Appendix H). We hope to use this approach, together with the topic-modelling work in Section 3.2 above, to integrate insights from WP3 and WP4 and enable models which link comments with the content of their article.

7 Associated outputs

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

Description	URL	Availability
Code for author profiling (Koloski et al., 2020a)	github.com/EMBEDDIA/PAN2020-Celebrity-Profiling	Public (MIT)
Code for topic modelling (Zosa et al., 2021)	github.com/EMBEDDIA/croatian-topic-api	Public (MIT)
Code for fake news spreader detection (Koloski et al., 2020b)	github.com/EMBEDDIA/PAN2020-Fake-News-Spreaders-Profiling	Public (MIT)
Code for neural fake news detection (Koloski et al., 2021)	github.com/EMBEDDIA/covid19-fake-news	Public (MIT)
Code for cross-lingual sentiment (Robnik-Šikonja et al., 2021)	github.com/EMBEDDIA/cross-lingual-classification-of-tweet-sentiment	Public (MIT)
Code for opinion detection (in progress)	github.com/EMBEDDIA/opinion-detection	To become public*
Code for thread reconstruction (in progress)	github.com/EMBEDDIA/threadStructure	To become public*

*Resources marked here as "To become public" are available only within the consortium while under development and/or associated with work yet to be published. They will be released publicly when the associated work is completed and published.

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

deliverable as appendices:

Citation	Status	Appendix
Koloski, B., Pollak, S., & Škrlj, B. (2020a). Know your neighbors: Efficient author profiling via follower tweets. In Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).	Published	Appendix A
Zosa, E., Shekhar, R., Karan, M., & Purver, M. (2021). Not all comments are equal: Insights into comment moderation from a topic-aware model.(Submitted, under review)	Submitted	Appendix B
Koloski, B., Pollak, S., & Škrlj, B. (2020b). Multilingual detection of fake news spreaders via sparse matrix factorization. In Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).	Published	Appendix C
Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of COVID-19 related fake news via neural stacking. In T. Chakraborty, K. Shu, H. Bernard, H. Liu, & M. Akhtar (Eds.), Combating online hostile posts in regional languages during emergency situation. CONSTRAINT 2021. (preprint available at arXiv:2101.03988).	Published	Appendix D
Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. Slovenščina 2.0, 9(1), 1–25. doi: 10.4312/slo2.0.2021.1.1-25	Published	Appendix E
Tabak, T., & Purver, M. (2020). Temporal mental health dynamics on social media. In Proceedings of the 1st workshop on NLP for COVID-19 (part 2) at EMNLP 2020. Association for Computational Linguistics.	Published	Appendix F
Shekhar, R., Karan, M., Purver, M., Pelicon, A., Pollak, S., Žagar, A. & Robnik Šikonja, M. (2021). Detecting and Explaining Viewpoints in Context. Task proposal, submitted to SemEval 2022.	Submitted	Appendix G
Rohanian, M., Hough, J., & Purver, M. (2020). Multi-modal fusion with gating using audio,lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. In Proceedings of INTERSPEECH. Shanghai, China: ISCA. doi: 10.21437/Interspeech.2020-2721	Published	Appendix H

Bibliography

Abbott, R., Ecker, B., Anand, P., & Walker, M. (2016, May). Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4445–4452). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L16-1704

Allaway, E., & McKeown, K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

Barker, E., & Gaizauskas, R. (2016). Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of the ACL 3rd Workshop on Argument Mining.*

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017). N-GrAM: New Groningen author-profiling model. In *Proceedings of the 8th International Conference of the CLEF Association (CLEF 2017).*

Batra, V., He, Y., & Vogiatzis, G. (2018). Neural caption generation for news images. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., ... Ostendorf, M. (2011). Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media* (p. 48-57). Association for Computational Linguistics.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bošnjak, M., & Karan, M. (2019, August). Data set for stance and sentiment analysis from user comments on Croatian news. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 50–55). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-3707 doi: 10.18653/v1/W19-3707

Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., ... Nissim, M. (2016). GronUP: Groningen user profiling. In *Proceedings of the 7th International Conference of the CLEF Association (CLEF 2016).*

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., ... Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In S. Renals & S. Bengio (Eds.), *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Revised Selected Papers* (Vol. 3869, pp. 28–39). Springer.

Celli, F., Riccardi, G., & Ghosh, A. (2014). CorEA: Italian news corpus with emotions and agreement. In *Conferenza di Linguistica Computazionale.*

Celli, F., Stepanov, E., Poesio, M., & Riccardi, G. (2016). Predicting Brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)* (pp. 110–118). Osaka, Japan.

Celli, F., Stepanov, E. A., & Riccardi, G. (2016). Tell me who you are, i'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blogs. In *Proceedings of the Workshop: Natural Language Processing meets Journalism.*

Concannon, S., & Purver, M. (in preparation). *Detecting patient opinion in social media [exact title anonymised for review].* (Draft, for submission to BMJ Quality & Safety)

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51–60). Baltimore, Maryland, USA: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W14-3207` doi: 10.3115/v1/W14-3207

Das, A., Chakraborty, T., Solorio, T., Gambäck, B., Aguilar, G., Kar, S., & Garrette, D. (2020). Semeval-2020 Task 9: Sentiment analysis for code-mixed social media text. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020).* Barcelona, Spain.

Dell'Orletta, F., & Nissim, M. (2018). Overview of the EVALITA 2018 cross-genre gender prediction (gxg) task. In *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA).* Turin, Italy.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186).

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. Retrieved from `https://www.aclweb.org/anthology/2020.tacl-1.29` doi: 10.1162/tacl_a_00325

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska Symposium on Motivation 1971* (Vol. 19). University of Nebraska Press.

Elsner, M., & Charniak, E. (2011, June). Disentangling chat with local coherence models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1179–1189). Portland, Oregon, USA: Association for Computational Linguistics.

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 483–488). Beijing, China: Association for Computational Linguistics.

Jwa, H., Oh, D., Park, K., Kang, J., & Lim, H. (2019). exBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). *Applied Sciences*, *9*(19), 40-62.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the ACL* (pp. 655–665).

Kiesling, S. F., Pavalanathan, U., Fitzpatrick, J., Han, X., & Eisenstein, J. (2018). Interactional stancetaking in online forums. *Computational Linguistics*, *44*(4), 683–718.

Koloski, B., Pollak, S., & Škrlj, B. (2020a). Know your neighbors: Efficient author profiling via follower tweets. In *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).*

Koloski, B., Pollak, S., & Škrlj, B. (2020b). Multilingual detection of fake news spreaders via sparse matrix factorization. In *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020).*

Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlj, B. (2021). Identification of COVID-19 related fake news via neural stacking. In T. Chakraborty, K. Shu, H. Bernard, H. Liu, & M. Akhtar (Eds.), *Combating online hostile posts in regional languages during emergency situation. CONSTRAINT 2021* (Vol. 1402). Springer.

Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, *53*(1), 1–37.

Kumar, S., & Carley, K. M. (2019). Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5047–5058).

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, *abs/1901.07291*. Retrieved from `http://arxiv.org/abs/1901.07291`

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Li, X., Bing, L., Lam, W., & Shi, B. (2018). Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 946–956). Melbourne, Australia: Association for Computational Linguistics.

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`

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, *14*(8), e0221152.

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2021). Topic modelling discourse dynamics in historical newspapers. In *Post-proceedings of the 5th conference digital humanities in the nordic countries (dhn 2020)* (pp. 63–77). Germany: Schloss Dagstuhl Leibniz Center for Informatics. (Digital Humanities in the Nordic Countries Conference, DHN ; Conference date: 21-10-2020 Through 23-10-2020)

Martinc, M., & Pollak, S. (2019). Pooled LSTM for Dutch cross-genre gender classification. In *Proceedings of the shared task on cross-genre gender prediction in Dutch at CLIN29 (GxG-CLIN29) co-located with the 29th conference on computational linguistics in the Netherlands (clin29).*

Martinc, M., Škrlj, B., & Pollak, S. (2019a). Fake or not: Distinguishing between bots, males and females. In *Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).*

Martinc, M., Škrlj, B., & Pollak, S. (2019b). Who is hot and who is not? profiling celebs on Twitter. In *Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).*

Miao, Y., Yu, L., & Blunsom, P. (2016). Neural variational inference for text processing. In *International conference on machine learning* (pp. 1727–1736).

Miura, Y., Taniguchi, T., Taniguchi, M., & Ohkuma, T. (2017). Author profiling with word+character neural attention network. In *Proceedings of the 8th International Conference of the CLEF Association (CLEF 2017).*

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018).* New Orleans, LA, USA.

Morgan, J. T., Oxley, M., Bender, E., Zhu, L., Gracheva, V., & Zachry, M. (2013). Are we there yet?: The development of a corpus annotated for social acts in multilingual online discourse. *Dialogue and Discourse*, *4*(2), 1-33.

Müller, S., Huonder, T., Deriu, J., & Cieliebak, M. (2017). Topicthunder at semeval-2017 task 4: Sentiment classification using a convolutional neural network with distant supervision. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 766–770). Vancouver, Canada: Association for Computational Linguistics.

Napoles, C., Tetreault, J., Pappu, A., Rosato, E., & Provenzale, B. (2017). Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop* (pp. 13–23).

Nasreen, S., Purver, M., & Hough, J. (2019a). A corpus study on questions, responses and misunderstanding signals in conversations with Alzheimer's patients. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers.* London, United Kingdom: SEMDIAL.

Nasreen, S., Purver, M., & Hough, J. (2019b). Interaction patterns in conversations with Alzheimer's patients. In *7th International Conference on Statistical Language and Speech Processing.* Ljubljana, Slovenia. (Abstract and poster)

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., & Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, *7*, e559. Retrieved from `https://doi.org/10.7717/peerj-cs.559` doi: 10.7717/peerj-cs.559

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 482–491). Avignon, France: Association for Computational Linguistics.

Ramisa, A., Yan, F., Moreno-Noguer, F., & Mikolajczyk, K. (2018). BreakingNews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(5), 1072-1085.

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.*

Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. In *Working Notes Papers of the CLEF.*

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF.*

Rangel, F., Rosso, P., y Gómez, M. M., Potthast, M., & Stein, B. (2018). Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in Twitter. In *Working Notes Papers of the CLEF.*

Riccardi, G., Bechet, F., Danieli, M., Favre, B., Gaizauskas, R., Kruschwitz, U., & Poesio, M. (2016). The SENSEI project: Making sense of human conversations. In J. Quesada & et al. (Eds.), *Future and Emerging Trends in Language Technology* (pp. 10–33).

Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2020). Cross-lingual transfer of Twitter sentiment models using a common vector space. In *Submitted.*

Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. *Slovenščina 2.0*, *9*(1), 1–25. doi: https://doi.org/10.4312/slo2.0.2021.1.1-25

Rohanian, M., Hough, J., & Purver, M. (2019). Detecting depression with word-level multimodal fusion. In *Proceedings of INTERSPEECH* (pp. 1443–1447). Graz, Austria: ISCA. (ISSN 1990-9772)

Rohanian, M., Hough, J., & Purver, M. (2020). Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. In *Proceedings of INTERSPEECH.* Shanghai, China: ISCA. (ISSN 1990-9772) doi: 10.21437/Interspeech.2020-2721

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 502–518). Vancouver, Canada: Association for Computational Linguistics.

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`

Schabus, D., Skowron, M., & Trapp, M. (2017). One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1241–1244).

Shao, C., Ciampaglia, G., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, *9*(1), 1–9.

Shekhar, R., Karan, M., Purver, M., Pelicon, A., Pollak, S., Žagar, A., & Šikonja, M. R. (2021). *Detecting and explaining viewpoints in context.* (Draft, submitted as a task proposal to SemEval-2022)

Shekhar, R., Pranjić, M., Pollak, S., Pelicon, A., & Purver, M. (2020). Automating news comment moderation with limited resources: Benchmarking in Croatian and Estonian. *Journal of Language Technology and Computational Linguistics*, *34*, 49-79. (Special Issue on Offensive Language)

Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., & Fernández, R. (2019). Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. 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. 2578–2587). Minneapolis, Minnesota: Association for Computational Linguistics.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial workshop on discourse and dialogue* (pp. 97–100). Cambridge, Massachusetts.

Shu, K., Bernard, H., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 43–65). Springer.

Tabak, T., & Purver, M. (2020). Temporal mental health dynamics on social media. In *Proceedings of the 1st workshop on NLP for COVID-19 (part 2) at EMNLP 2020.* Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.7

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., & Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. In *Proceedings of the 9th International Conference of the CLEF Association (CLEF 2018).*

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Proceedings of Text, Speech, and Dialogue, TSD 2020.* doi: 10.1007/978-3-030-58323-1_11

Vamvas, J., & Sennrich, R. (2020). X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Verhoeven, B., Daelemans, W., & Plank, B. (2016). TwiSty: a multi-lingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC).* ELRA.

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., & King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of LREC* (p. 812-817).

Wiegreffe, S., & Marasović, A. (2021). Teach me to explain: A review of datasets for explainable nlp..

Zosa, E., & Granroth-Wilding, M. (2019). Multilingual dynamic topic model. In *Proceedings of the international conference on recent advances in natural language processing (ranlp 2019)* (pp. 1388–1396). Varna, Bulgaria: INCOMA Ltd. doi: 10.26615/978-954-452-056-4_159

Zosa, E., Shekhar, R., Karan, M., & Purver, M. (2021). *Not all comments are equal: Insights into comment moderation from a topic-aware model.* (Submitted, under review)

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2016). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2438–2448). Osaka, Japan: The COLING 2016 Organizing Committee.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., ... Augenstein, I. (2018). Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, *54*(2), 273-290. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0306457317303746` doi: https://doi.org/10.1016/j.ipm.2017.11.009

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., & Pollak, S. (2021). tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*, *65*, 101104. doi: https://doi.org/10.1016/j.csl.2020.101104

Appendix A: Know your Neighbors: Efficient Author Profiling via Follower Tweets

Know your Neighbors: Efficient Author Profiling via Follower Tweets

Notebook for PAN at CLEF 2020

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

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

Abstract User profiling based on social media data is becoming an increasingly relevant task with applications in advertising, forensics, literary studies and sociolinguistic research. Even though profiling of users based on their textual data is possible, social media such as Twitter offer also insight into the data of a given user's followers. The purpose of this work was to explore how such follower data can be used for profiling a given user, what are its limitations and whether performances, similar to the ones observed when considering a given user's data directly can be achieved. In this work we present our approach, capable of extracting various feature types and, via sparse matrix factorization, learn a dense, low-dimensional representations of individual persons solely from their followers' tweet streams. The proposed approach scored second in the PAN 2020 Celebrity profiling shared task, and is computationally non-demanding.

1 Introduction

User profiling on social media is becoming an increasingly relevant task when detecting problematic users or bots. In the era of social media, text-based representations of such users need to be learned, which is becoming a lively research area [5]. Online social media, such as Twitter, offer an unique opportunity to test to what extent properties of users can be predicted, and what potential implications of such learning endeavours are [3]. This paper discusses the challenge of predicting a given user's property based *solely* on the information captured from a given *user's followers' texts*. The paper explores to what extent the follower data offers profiling capabilities and what are its limitations. The schematic overview of the scenario considered in this work is shown in Figure 1. The remainder of this work is structured as follows. In Section 2 we present the related work, followed by the description of the proposed system (Section 4), experimental evaluation (Section 6) and the concluding remarks in Section 8.

2 Related Work

One of the first author profiling tasks was gender prediction by Koppel et al. [4], who conducted experiments on a subset of the British National Corpus and found that women have a more relational writing style and men have a more informational writing

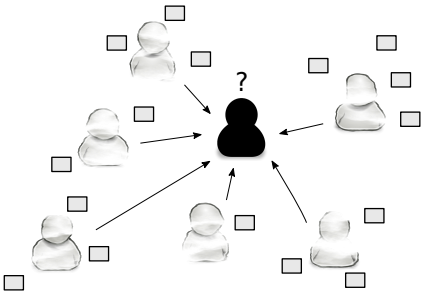


Figure 1: Schematic overview of the considered task. The gray boxes around the users surrounding the user of interest (middle node), are the users' tweets.

style. While deep learning approaches have been recently prevailing in many natural language processing and text mining tasks, the state-of-the-art research on gender classification mostly relies on extensive feature engineering and traditional classifiers.

Examples of previous PAN competition winners include [2] (who used support vector machines), however, the second ranked solution [7] was even simpler, employing only logistic regression classifier with features containing also emoji information and similar. In PAN 2016, the best gender classification performance was achieved by [8], who employed a Logistic regression classifer and used word unigrams, word bigrams and character four-gram features.

PAN 2016 AP shared task also dealt with age classification. The winners in this task [12] used a linear SVM model and employed a variety of features: word, character and POS tag n-grams, capitalization (of words and sentences), punctuation (final and per sentence), word and text length, vocabulary richness, emoticons and topic-related words. We acknowledge also the research of [1], who among other classification tasks also dealt with the prediction of text author's occupation on Spanish tweets. They evaluated several classification approaches (bag of terms, second order attributes representation, convolutional neural network and an ensemble of n-grams at word and character level) and showed that the highest performance can be achieved with an ensemble of word and character n-grams. Finally, the modeling task addressed in this work is similar to the last year's PAN Celebrity Profiling Challenge that aimed at predicting age, gender, fame and occupation[13], from which we also sourced some of the ideas used in the final models. The winning approach last year used tf-idf features with logistic regression and SVM classifiers [10].

3 Dataset Description and Preprocessing

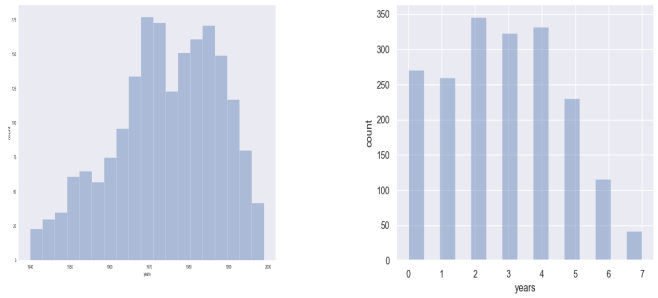
The training set for the PAN 2020 Celebrity Profiling shared task is composed of English tweets of follower feeds of 1,920 celebrities, labeled in three categories: gender,

occupation and birthyear. The dataset is balanced towards gender and occupation, while the birthyear label is not balanced. Distribution of the gender and occupation data is shown in Table 1 and birthyear data is presented in Figure 2 containing the original distribution and the augmented one, as described in Section 6.

Table 1: Distribution of gender and occupation labels.

Gender	Count
male	1072
female	1848

Occupation	Count
sports	480
performer	480
creator	480
politics	480



(a) Birthyears from the initial dataset

(b) Birthyears after augmenting the dataset

Figure 2: Overview of the birthyear distributions.

For getting the data prepared we firstly select 20 tweets for 10 authors for each celebrity, meaning 200 tweets in total for each celebrity in our data. Next, the tweet data is concatenated and preprocessed, as discussed next.

4 Feature Construction and Classification Model

The following section includes description of the proposed method and its intermediary steps.

Before feature construction, dimensionality reduction and classifier application, in the initial step we construct multiple representations of a given user that we denote as a collection C. The space of constructed features, similarly to [6] and [7], is based on:

– original text
– punctuation free - from the original text we removed punctuation
– stop-words free - from the punctuation free version stop words are removed

5 Authomatic feature seleciton

The collection C consists of multiple representations for each author, offering large space of potential features. We focused on character and word-level features to capture potentially interesting semantics. For this step, we used the SciKit-learn's [9] word tokenizer. The generated features are described as follows:

– character based - from each part in the collection C we generate character n-grams (up to 1 or 2 characters) and up to $\frac{n}{2}$ maximum allowed character features.
– word based- from each part in the collection C we generate word n-grams (up to 1,2,3 words) and up to $\frac{n}{2}$ maximum allowed word n-gram features

At the conclusion of the pipeline execution, we have prepared word and character features from each celebrity's collection of tweets, ready to be used in the feature selection step, which are finally joined via SciKit-learn's FeatureUnion.

5.1 Dimensionality reduction via matrix factorization

Finally, we perform sparse singular value decomposition (SVD)[1] that can be summarized via the following expression:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T.$$

The final representation (embedding) $\boldsymbol{E}$ is obtained by multiplying back only a portion of the diagonal matrix ($\boldsymbol{\Sigma}$) and $\boldsymbol{U}$, giving a low-dimensional, compact representation of the initial high dimensional matrix. Note that $\boldsymbol{E} \in \mathbb{R}^{|D| \times d}$, where d is the number of diagonal entries considered. The obtained $\boldsymbol{E}$ is suitable for a given down-stream learning task, such as classification (considered in this work). Note that performing SVD in the text mining domain is also commonly associated with the notion of *latent semantic analysis*.

5.2 Classifier selection

For each sub task we performed extensive grid-search using [9] GridSearchCV and found classifiers that suited task the most. Following this goal we conducted a series of experiments, consisting of trying different environments and linear models as presented in the Section 6. Among the one we used were (SciKit learn's [9]) Support Vector Machines, Random Forests and Logistic Regression.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

6 Experiments

Series of experiments were executed in order to find the best embedding space and model. We explored various ways of modeling the birthyear variable:

- **R** - regression - where we applied linear regression and XGBoost Regressor [9] learner to derive a simple model to predict the years, where we predicted birthyear in the interval:

$$\max(1949, \min(\text{predicted\_year}, 1999)).$$

- **FC** - full classification - we applied classification learner to the task discrimination between each of the 60 classes (one year = one class)
- **AC** - altered classification - we applied classification to an altered label space where we reduced the number of labels to more balanced intervals, finally obtaining 8 of them, hence: 1949 - 1958, 1959 - 1966, 1967 - 1973, 1974 - 1980, 1981 - 1986, 1987 - 1991, 1992 - 1995, 1996 - 1999. For the final reverse prediction in the interval back we used the following estimates.
  1. predicting the middle of the interval
  2. predicting random year from the interval

For all tasks we considered GridSearchCV over parameter space to find best hyper-parameter configuration, dimension number k and the number of features to be generated n. By doing 10-fold cross validation, the grid consisted of reducing the dimensions parametrized by k in the following interval:

$$k \in [128, 256, 512, 640, 768, 1024, 2048]$$

and the number of generated n features from the interval

$$n \in [2500, 5000, 10000, 20000, 30000, 50000].$$

The initial dataset was split to training(90%) and evaluation(10%) sets from after which we obtain C_{training} and $C_{\text{evaluation}}$. Once constructed, the feature space was considered for learning. We experimented with XGBoost, logistic regression and linear SVMs, of which hyperparameters we optimized in 5 fold cross validation. Finally, we tested the performance on the $C_{\text{evaluation}}$ set.

7 Results

This section includes the results of the empirical evaluation, used to select the final model. The obtained results are shown in table 2.

Table 2: Final evaluation on training data on TIRA.

name	#features	#dimensions	f1 age	f1 gender	f1 occupation	crank
model-**AC**-2	20000	512	**0.358**	**0.665**	0.656	**0.516**
model-**AC**-1	20000	512	0.346	0.663	**0.669**	0.509
model-**FC**-2	10000	512	0.313	0.639	0.632	0.473
model-**FC**-1	10000	512	0.291	0.605	0.648	0.452
model-**R**	10000	512	0.298	0.612	0.613	0.453
baseline-ngrams	#	#	0.362	0.584	0.521	0.469

The best scoring model is model-**AC**-2, which we chose for (final) test evaluation. Its hyperparameters were: $n = 20000$ features reduced to $k = 512$, while the Logistic Regression (occupation and age)'s regularization was set to $\lambda_2 = 1$. For gender, the SVM's hyperparameters were $\lambda_2 = 1$, gamma factor = scale and the polynomial kernel was used.

The best preforming model of experiments conducted in Section 6 yielded the following results on the test set on the TIRA site. We next present the official ranking of the proposed solution on the final TIRA test set.

TEAM	TEST-DATASET			
	CRANK	AGE	GENDER	OCCUPATION
baseline-ngram-celebrity-tweets	0.631	0.500	0.753	0.700
hodge20	0.577	0.432	0.681	0.707
koloski20	0.521	0.407	0.616	0.597
tuksa20	0.477	0.315	0.696	0.598
baseline-ngram-follower-tweets	0.469	0.362	0.584	0.521
random	0.333	0.333	0.500	0.250

Figure 3: The proposed submission achieved 2nd place (koloski20) (not accounting for full-tweet baseline).

The proposed system scored the second highest (the first listed in Figure 3 is the baseline based solely on a given author's tweet stream. It outperforms the generic baselines, whilst maintaining a lower dimension of the representation.

8 Discussion and conclusions

As not a single competing submission (Figure 3) achieved performance above the baseline trained on a given person's tweets, this task demonstrates that such type of classification is exceptionally hard, and needs to be fundamentally re-thought to overcome the full-information models aware of a given person's tweets. Significant improvement was achieved from the thresholding of the years and reducing the number of age classes to less than initial given, since the f1-score of age was based on the hit interval for years, giving us an uphold for varying different interval pooling strategies, namely we used two: first one based on generating the middle year in our predefined year interval and

the second was guessing a random number from the interval. The celebrity's own tweets and tweets of its followers gave competitive f1-scores while using relatively simple features (no emojis or similar) and computationally efficient methods representation construction methods. Finally, the score was calculated by calculating the harmonic mean of f1-scores:

$$cRank = \frac{1}{\frac{1}{f1\_occupation} + \frac{1}{f1\_birthyear} + \frac{1}{f1\_gender}}$$

As seen in the 7 section we believe that improving one the score on one subtask will only benefit the whole model if we keep or improve the scores on the other subtasks.

Further work will include trying out different division of the birthyear values by trying out different thresholds, possibly trying to inject more semantically enriched vectorization features [11] of tweets or improve the way the data is polled to build the data representation for a single celebrity.

9 Acknowledgements

The work of the last author was funded by the Slovenian Research Agency through a young researcher grant. The work was also supported by the Slovenian Research Agency (ARRS) core research programme *Knowledge Technologies* (P2-0103), an ARRS funded research project *Semantic Data Mining for Linked Open Data* (financed under the ERC Complementary Scheme, N2-0078) and EU 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

1. Aragón, M.E., López-Monroy, A.P.: Author profiling and aggressiveness detection in spanish tweets: Mex-a3t 2018. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)
3. Batool, R., Khattak, A.M., Maqbool, J., Lee, S.: Precise tweet classification and sentiment analysis. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). pp. 461–466. IEEE (2013)
4. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
5. Markov, I., Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Gelbukh, A.: Author profiling with doc2vec neural network-based document embeddings. In: Mexican International Conference on Artificial Intelligence. pp. 117–131. Springer (2016)
6. Martinc, M., Blaž Škrlj Pollak, S.: Fake or not: Distinguishing between bots, males and. CLEF 2019 Evaluation Labs and Workshop – Working Notes Papers (2019)
7. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers (2017)

8. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
10. Radivchev, V., Nikolov, A., Lambova, A.: Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/
11. Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S.: tax2vec: Constructing interpretable features from taxonomies for short text classification. arXiv preprint arXiv:1902.00438 (2019)
12. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: Gronup: Groningen user profiling notebook for PAN at clef 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
13. Wiegmann, M., Stein, B., Potthast, M.: Celebrity profiling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2611–2618 (2019)

Appendix B: Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model

Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model

Anonymous IWCS submission
The style files are borrowed from ACL-IJCNLP 2021

Abstract

Moderation of reader comments is a significant problem for online news platforms. Here, we experiment with models for automatic moderation, using a dataset of comments from a popular Croatian newspaper. Our analysis shows that while comments that violate the moderation rules mostly share common linguistic and thematic features, their content varies across the different sections of the newspaper. We therefore propose to make our models topic-aware, incorporating semantic features from a topic model into the classification decision. Our results show that topic information improves the performance of the model, increases its confidence in correct outputs, and helps us understand the model's outputs.

1 Introduction

Most newspapers publish their articles online, and allow readers to comment on those articles. This can increase user engagement and page views, and provides readers with an important route to public freedom of expression and opinion, with the ability to interact and discuss with others. Comment sections usually provide some degree of anonymity;[1] while improving access, this can also encourage inappropriate behaviour, and publishers therefore usually employ some moderation policy to regulate content and to ensure legal compliance (in some cases, publishers can be held responsible for user-contributed content on their sites).

One possible approach is a 'moderate then publish' policy, in which comments must be approved by a moderator before they appear; this requires significant manpower and introduces delays and limitations into the user conversation (for example, the New York Times only allows comments for one day after article publication[2]). On the other hand, a 'publish then moderate' strategy, in which comments are published immediately, and later removed if necessary, is less effective at blocking toxic or illegal content. Combined with the increase in comment volumes in recent years (Shekhar et al., 2020, found a 250% increase in comments per article from 2009 to 2015) there is increasing interest in automatic moderation methods (see e.g. Pavlopoulos et al., 2017a), either as stand-alone tools or for integration into human moderators' practices (Schabus and Skowron, 2018).

Detecting comments that need moderators' attention is usually approached as a text classification task (see e.g. Pavlopoulos et al., 2017a); but comments can be blocked for a range of reasons (Shekhar et al., 2020). One is the presence of offensive language, a well-studied NLP task (see Section 2 below); however, others include advertising or spam, illegal content, spreading misinformation, trolling and incitement — all distinct categories which might be expected to show distinct features, and perhaps to vary according to the content being commented on. Another aspect that distinguishes the comment moderation task from the usual text classification tasks in NLP is the need for interpretable or explainable models: if classifiers are to be used by human moderators within publishers' working practices, they must be able to understand the outputs (Švec et al., 2018).

Here, we therefore investigate models which can provide both an aspect of interpretability and the ability to take account of the topics being discussed, by incorporating topic distribution information into the comment classifier. Specifically, we incorporate semantic representations learned by the Embedded Topic Model (ETM) (Dieng et al., 2020) into a classifier pipeline based on Long Short-Term Mem-

[1] Some newspapers allow completely anonymous posting; some require commenters to create an account with a username, but this does not usually reveal their true identity.

[2] NYT Comment FAQ: https://nyti.ms/2PF02kj

ory (LSTM) networks (Hochreiter and Schmidhuber, 1997). Our model improves performance by 4.4% over a text-only approach on the same dataset (Shekhar et al., 2020), and is more confident in the correct decisions it makes. Inspection of the topic distributions then allows us to interpret the source of these improvements, and reveals how different newspaper sections have different language and topic distributions, including differences in the kind of comments that need moderation.[3]

2 Related Work

Automated news comment moderation Most research on this task so far formulates it as a text classification problem: for a given comment, the model must predict whether the comment violates the newspaper's policy. Approaches to classification vary, however. Nobata et al. (2016) use a range of linguistic features, e.g. lexicon and n-grams. Schabus et al. (2017) use a bag-of-words approach. Pavlopoulos et al. (2017a) and Švec et al. (2018) use neural networks, specifically RNNs with an attention mechanism. Recently, Tan et al. (2020) and Tran et al. (2020) apply a modified BERT model (Devlin et al., 2019).

Some approaches go beyond the comment text itself: Gao and Huang (2017) add information like user ID and article headline into their RNN to make the model context-aware; Pavlopoulos et al. (2017b) incorporate user embeddings; Schabus and Skowron (2018) incorporate the news category metadata of the article to improve the performance; and Risch and Krestel (2018) add both user and article features into the model. However, no work so far investigates automatic modelling of topic (rather than relying on categorical metadata), or applies this to the comments rather than just their parent articles.

Some steps towards model intepretability and output explanation have also been taken: both Švec et al. (2018) and Pavlopoulos et al. (2017a) use an attention saliency map to highlight the possibly problematic words. However, we are not aware of any work using higher-level topic information as a route to understanding model outputs.

Available datasets Several datasets have been created for the news comment moderation task. Nobata et al. (2016) provide 1.43M comments posted on Yahoo! Finance and News over 1.5 years, in which 7% of the comments are labelled as abusive via a community moderation process. Gao and Huang (2017) contains 1.5k comments from Fox News, annotated with specific hateful/non-hateful labels as a post-hoc task, and having 28% hateful comments. However, both are relatively small, and their labelling methods mean that neither dataset is entirely representative of the moderation process performed by newspapers.

Pavlopoulos et al. (2017a) provides 1.6M comments from Gazzetta, a Greek sports news portal, over c.1.5 years. Here, 34% of comments are labelled as blocked, and the labels are derived from the newspapers human moderators and journalists. Schabus et al. (2017) and Schabus and Skowron (2018) provide a dataset from a German-language Austrian newspaper with 1M comments posted over 1 year, out of which 11,773 comments are annotated using seven different rules. Again, this annotation was performed by the newspaper's moderators using their policy; but again, the dataset is relatively small. Švec et al. (2018) use a larger dataset of 20M comments from a Slovak newspaper, annotated for insults, racism, profanity and spam - but do not make it publicly available.

More recently, Shekhar et al. (2020) present a dataset from 24sata, Croatia's most widely read newspaper.[4] This dataset is significantly larger (10 years, c.20M comments); and moderator labels include not only a label for blocked comments, but a record of the reason for the decision according to a 9-class moderation policy. However, their experiments show that classifier performance is limited, and transfers poorly across years. Here, we therefore use this dataset (see Section 4), with a view to improving performance and applying a topic-aware model to improve and better understand the robustness in the face of changing topics.

Related tasks More attention has been given to related tasks, most prominently the detection of offensive language, hate speech and toxicity. For this task, datasets have been sourced from many online platforms, e.g. Twitter (Davidson et al., 2017), Facebook (Ljubešić et al., 2019), YouTube (Obadimu et al., 2019) and Reddit (Qian et al., 2019). The exact task definition and focus vary: for example, Waseem and Hovy (2016) and Waseem (2016) annotate for racism and sexism, while Davidson et al. (2017) look at more general offensive and hate speech categories.

[3] Source code will be publicly available upon acceptance.

[4] http://24sata.hr/

Multiple shared tasks (e.g. OffensEval, Zampieri et al., 2019, 2020) have been proposed to measure progress in the moderation process; and recently, focus has shifted from English to other languages too, e.g. EVALITA 2018 for Italian (Bai et al., 2018), GermEval 2018 for German (Wiegand et al., 2018). A comprehensive survey of dataset collection is provided by Poletto et al. (2020) and Vidgen and Derczynski (2020).[5]

Topic Modelling Topic models capture the themes from a collection of documents through the co-occurence statistics of the words used in a document. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a popular method for capturing these themes (also known as *topics*), is a generative document model where a document is a mixture of topics expressed as a probability distribution over the topics and a topic is a distribution over the words in a vocabulary.

The Embedded Topic Model (ETM) (Dieng et al., 2020) is an LDA-like topic modelling method that exploits the semantic information captured in word embeddings during topic inference. The advantage of ETM over LDA is that it combines the advantages of word embeddings with topic modelling and has been shown to produce more coherent topics than regular LDA.

3 Proposed Model

Our aim is to incorporate document-level semantics with textual features in the comment moderation process. To this end, we came up with several model architectures that combine a language model with topic features.

For the comment text representation, we use a bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997). For a given comment, the text is passed through an embedding layer then a BiLSTM where the output of the final hidden state is taken as the encoded representation of the comment.

For the topic features, we use topics from a trained **Embedded Topic Model (ETM)** (Dieng et al., 2020). In the ETM, the topic-term distribution for topic k, β_k, is induced by a matrix of word embeddings ρ and the topic embedding α_k which is a point in the embedding space:

$$\beta_k = softmax(\rho^T \alpha_k) \quad (1)$$

[5] http://hatespeechdata.com/ provides a comprehensive list of relevant datasets.

The topic embeddings, α, are learned during topic inference while the word embeddings ρ can be pretrained or also learned during topic inference. In this work, we use pretrained embeddings.

The document-topic distribution of a document d, θ_d, is drawn from the logistic normal distribution (LN) whose mean and variance come from an inference network:

$$\theta_d \sim LN(\mu_d, \sigma_d) \quad (2)$$

Given a trained ETM, we infer the θ_d of an unseen document d which we take as our **document-topic vector (DTV)**. Then we compute the **document-topic embedding (DTE)** as the weighted sum of the embeddings of the topics in doc d, where the weight corresponds to the probability of the topic in that document:

$$DTE = \sum_{k=0}^{K} \alpha_k \theta_{d,k} \quad (3)$$

where α_k is the topic embedding of topic k, and $\theta_{d,k}$ is the probability of topic k in doc d.

We propose two fusion mechanisms to combine the comment text and topic representation: *early* and *late* fusion. In early fusion, topic features are concatenated with the comment word embeddings and then passed to the BiLSTM. In **EarlyFusion1 (EF1)**, only DTV is concatenated with the word embeddings; **EarlyFusion2 (EF2)** uses DTE instead of DTV; and **EarlyFusion3 (EF3)** uses both DTE and DTV. In late fusion, topic features are concatenated with the output representations of the BiLSTM, and passed to the MLP for classification. Again, **LateFusion1 (LF1)** uses DTV; **LateFusion2 (LF2)** uses DTE; and **LateFusion3 (LF3)** uses both. Figure 1 shows the architecture.

Note that the late fusion architecture is similar to the Neural Composite Language Model (NCLM, Chaudhary et al., 2020), which also combines topic information with a language model. Unlike NCLM, our model does not do joint training of the topic model and language model. Instead, we train the topic model, extract topic features and then use them to train the BiLSTM-based classifier. Another difference is that since we use the ETM as our topic model, we use the topic embeddings to obtain a dense topic representation which we call the DTE. In the NCLM, the equivalent of a DTE is the Explainable Topic Representation (ETR), which is computed as the sum of the word embeddings of

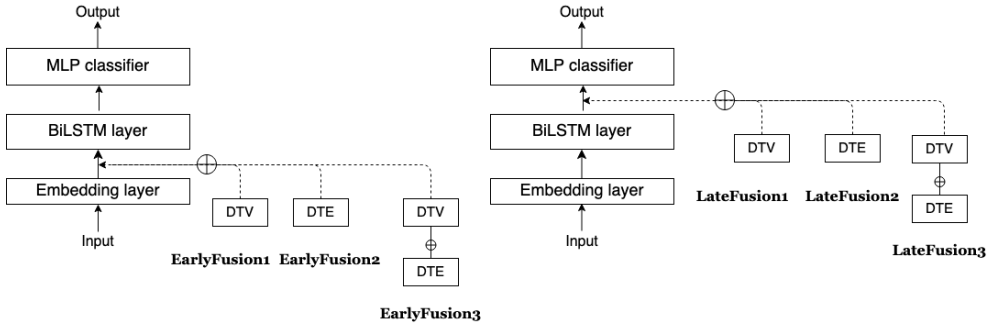


Figure 1: Proposed model architectures combining text and topic features.

Comment Moderation Data			
	Blocked	Non-blocked	Blocking Rate
Train	4984	75016	6.23%
Valid	642	9358	6.42%
Test	37271	438142	7.84%
Topic Modelling Data			
	Blocked	Non-blocked	Blocking Rate
Train	34863	36725	48.70%
Valid	4880	5120	48.80%

Table 1: Details of datasets used experiments.

the top N terms in the document, where N is a hyperparameter.

4 Dataset

We use the 24sata comment dataset (Shekhar et al., 2020), introduced in Section 2. This contains c.21M comments on 476K articles from the years 2007-2019, mostly in the Croatian language. There is no filtering involved in the data selection, which reduces the selection bias. The dataset has details of comments blocked by the 24sata moderators, based on a set of moderation rules, vary from hate speech to abuse to spam (see Shekhar et al., 2020, for rule description).[6] The dataset also identifies the article under which a comment was posted, together with the section/sub-section of the newspaper the article appeared in. These sections/sub-sections relate to the content of the article: For example, the Sport section contains sports-related news while the Kolumne (*Columns*) section contains opinion pieces. The largest section, Vijesti (*News*), is further subdivided as shown in Table 2.

Data selection We use data from 2018 for training and validation and 2019 data for testing. This reflects the realistic scenario where we use data collected from the past to make predictions. For training and validation, we randomly selected 50,000 articles out of 65,989 articles from 2018, sampling from the nine most-representative sections/sub-sections (Table 2). Each article comes with c.50 comments on average. We then randomly sample 90,000 comments from those articles as the training and validation data for our proposed models, thus retaining the naturally-occurring balance between blocked and non-blocked comments (only 6-8% of comments are marked as blocked).

To train ETM, we then sample another 90,000 comments with a roughly equal split between blocked and non-blocked comments. This is to encourage a diverse mix of topics from both comment classes. We also remove comments with less than 10 words for the ETM training data.

For the test set, we use all 475,413 comments associated with the 17,953 articles from 2019. Table 1 (upper part) provides the dataset details, with comment moderation blocking rate. For the test set, Table 2 provides details on the section and sub-section of the related articles. These top 9 sections account for more than 95% of the comments of the entire test set.

5 Experimental Setup

Baseline models As baselines, we use the following models trained only on text *or* topics:

- **Text only**: BiLSTM model with the comment text alone as input. The embedding layer is initialized with pretrained embeddings.
- **Document-topic vectors (DTV)**: MLP classifier with document-topic vectors as input.
- **Document-topic embedding (DTE)**: MLP classifier with document-topic embeddings.

[6] Rules are reproduced in the supplementary material.

Section (– Subsection)	Blocked	Non-blocked	Blocking Rate
Kolumne (*Columns*)	655	6382	9.31%
Lifestyle	2426	30985	7.26%
Show	6827	58896	10.39%
Sport	5882	80820	6.78%
Tech	382	7173	5.06%
Vijesti (*News*)	20094	239835	7.73%
– Crna kronika (*Crime*)	5917	62471	8.65%
– Hrvatska (*Croatia*)	3527	45170	7.70%
– Politika (*Politics*)	6088	80264	7.05%
– Svijet (*World*)	2625	31459	7.24%

Table 2: Details per section, and (for section Vijesti) sub-section, of the comment moderation test set.

- **DTV+E**: MLP classifier with concatenated document-topic vectors and embeddings.

Hyperparameters We use 300D word2vec embeddings, pre-trained on the Croatian Web Corpus (HrWAC, Ljubešić and Erjavec, 2011; Šnajder, 2014), for training the ETM and to initialize the embedding layer of the BiLSTM. The ETM is trained for 500 epochs for 100 topics with default hyperparameters from the original implementation. The BiLSTM is composed of one hidden layer of size 128 with dropout set to 0.5. We limit the comment length to the first 200 words. The MLP classifier is composed of one fully-connected layer, one hidden layer of size 64, a ReLU activation, and a sigmoid for classification with the classification threshold set to 0.5. We train all models for 20 epochs with early stopping based on the loss in the validation set.

6 Results

In Table 3, we present the performance of baseline and proposed models, measured as macro F1-score. All models combining text and topics perform better than the models that use only text or topic information. Surprisingly, the DTV model performs comparatively better than the DTE and DTV+E models, and performs almost as well as the text-only model; however, we show in Section 8 below that DTV is much less confident in its predictions than the text-only model. Overall, the best performing model is LF1, which improves the text-only model's performance by $+4.4\%$ (67.37% vs 62.97%); and improves by a similar amount over (Shekhar et al., 2020)'s results using mBERT (macro-F1 score 62.07 for year 2019).

Interestingly, we see wide variation in performance across news sections. We observe that Lifestyle and Tech are the easiest sections (best F1 over 0.72) while Politika (*Politics*) is the most difficult (best F1 below 0.62). The main cause appears to be that Lifestyle and Tech have the highest proportion of spam comments: on average, 49.44% of blocked comments in the test set are spam, but for Lifestyle and Tech this number rises to 77.25% and 69.63%, respectively.[7] As for the Politics section, we hypothesise that, excluding spam, the topics discussed in blocked and non-blocked comments have high overlap (see Section 7.2).

7 Analysis

This section presents a range of analyses that aim to shed light on the language and topics of blocked and non-blocked comments. We also analyze how language and topics vary across different news sections. Analysis is performed using the test data.

7.1 Content analysis

We analyze the linguistic differences between blocked and non-blocked comments and across different sections. First, we compare comment length between blocked and non-blocked comments; but as we can see from Table 4, blocked and non-blocked comments have similar mean length. If we further divide blocked comments into two subgroups, though — spam and non-spam — we find that on average, spam comments are longer than other comments. We observe a similar pattern across different sections (see Supp. material for the corresponding table).

Next, we ask whether there is any difference in the language used between blocked and non-blocked comments? To do this we used both unigram and bigram information. First. we measure the lexical diversity of the comments, measured using the mean-segmental type-token ratio (MSTTR). The MSTTR is computed as the mean of type-token ratio for every 1000 tokens in a dataset (van Miltenburg et al., 2018) to control for the dataset size. From Table 4, we see that non-blocked comments have higher MSTTR compared to blocked comments (0.62 vs 0.46), which suggests that non-blocked comments have a higher lexical diversity than blocked comments. However, when we divide blocked comments into spam and non-spam, we observe that non-spam blocked comments have a similar MSTTR to non-blocked comments (0.61 vs 0.62). The spam comments have lower MSTTR

[7] See Supplement for rule breakdown of the test data.

Section – Subsection	Text only	Topics only DTV	DTE	DTV+E	Text+Topic Combinations EF1	EF2	EF3	LF1	LF2	LF3
All	62.97	62.20	59.3	58.33	66.33	66.58	65.61	**67.37**	66.22	66.95
Kolumne	59.86	59.65	56.25	55.33	62.40	62.90	63.13	63.25	62.38	**63.6**
Lifestyle	69.21	70.07	65.93	64.47	72.73	70.9	69.36	72.00	72.39	**72.92**
Show	61.97	61.30	58.62	57.60	65.24	65.63	64.26	**66.50**	65.00	65.86
Sport	63.22	61.42	58.61	57.90	67.11	67.86	66.74	**68.26**	67.14	67.82
Tech	64.87	66.37	63.17	62.55	67.72	68.74	67.65	68.76	67.68	**69.15**
Vijesti (*News*)	62.38	61.49	58.79	57.77	65.58	65.99	65.24	**66.77**	65.53	66.24
– Crna kronika	64.67	63.98	61.03	59.84	68.10	68.88	68.11	**69.60**	67.89	68.88
– Hrvatska	63.61	63.50	60.10	58.93	67.24	66.86	65.95	67.90	67.12	**67.95**
– Politika	57.93	56.49	54.95	54.20	60.51	61.52	60.84	**61.61**	60.63	61.30
– Svijet	63.58	62.55	59.62	58.35	66.83	66.95	66.33	68.44	67.21	67.57

Table 3: Classifier performance measured as macro-F1.

(0.35 vs 0.61). This suggests that blocked comments (excluding spam) have as rich a vocabulary as non-blocked. Again, we see a similar pattern across different news sections (see Supp. material).

	Mean length	MSTTR
All	23.06	0.61
Non-blocked	23.01	0.62
Blocked	23.65	0.46
Blocked (non-spam)	19.16	0.61
Blocked (Spam only)	28.23	0.35

Table 4: Mean-segmental TTR and average length of comments

Next, we look at the top bigrams of blocked and non-blocked comments. For both classes, we collect all bigrams that occur at least 50 times and then rank those bigrams according to their pointwise mutual information (PMI) score. In general, we do not see many overlaps between the top bigrams of blocked and non-blocked comments across the different sections. Bigrams in the blocked comments indicate spam messages such 'iskustva potrebnog' (*experience required*), 'redoviti student' (*full-time student*) and 'prilika pruila' (*opportunity given*). Removing spam comments, we encounter bigrams used for swearing such as 'pas mater' (*damn it*) and 'jedi govna' (*eat sh*t*). In the non-blocked comments, the top bigrams are more relevant to the section they appear in. For instance, in the Vijesti section, top bigrams include 'new york', 'porezni obveznici' (*taxpayers*) and 'naftna polja' (*oil fields*) while in Sports, top bigrams include 'all star', 'grand slam' and 'man utd' and Lifestyle tends towards a more positive tone such as 'ugodan ostatak' (*pleasant rest*) and 'laku no' (*good night*). What this suggests is that the content of blocked comments tend to share commonalities across sections more than non-blocked comments; but again, these commonalities may be mostly within the spam category.

7.2 Topic analysis

Now we analyze how the topic distributions differ between blocked and non-blocked comments and across the sections. Our aim is to understand what subjects are discussed in these two comment classes and across the different sections, to gain insight into what characterises a blocked comment and a non-blocked one, and whether this varies between different sections.

We take the top topics of a document set by taking the mean of the topic distributions of all the documents in that set and ranking the topics according to their probability in this mean distribution. In this analysis, the document sets are the blocked and non-blocked comments. We take the top 15 topics for analysis because this is the average number of topics used by the comments (by this we mean the number of topics in a comment with a probability greater than zero).

For the entire test data, the top topics of non-blocked comments cover a diverse range of subjects from politics to football to scientific research (Figure 2). The top topics in blocked comments are dominated by spam and insults. Table 5 shows some of these topics (labels are manually assigned by native speaker). We provide the full topic list and descriptions in the Supplement. In Figure 2 we also see many topics shared between blocked and non-blocked comments.

We illustrate how different topics intersect between blocked and non-blocked comments across

Croatian football	dinamo, hajduk, zagreb, zagrebu, placu, europi, zagreba *(dynamo, hajduk, zagreb, zagreb, market, europe, zagreb)*
State and government	drava, drave, dravi, vlasti, dravu, vlade, vlada *(state, states, state, authorities, state, governments, government)*
Moderately offensive	gluposti, sramota, sram, glup, jadni, jadan, jadno, budale *(nonsense, shame, disgrace, stupid, miserable, miserable, miserable, fools)*
Death and illness	ena, ene, ljudi, osoba, osobe, enu, smrt, ovjeka *(woman, women, people, person, persons, woman, death, human)*
Civil war	srbi, hrvata, tito, srba, srbije, srbiji, srbima, srbija *(serbs, croats, tito, serbs, serbia, serbia, serbs, serbia)*

Table 5: Selected topics with English translations. The first two topics are prevalent in non-blocked comments, the next two are prevalent in blocked comments and the last is prevalent in both classes.

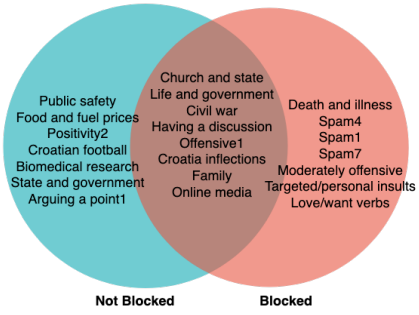


Figure 2: Top topics of the blocked and non-blocked comments for the entire test set.

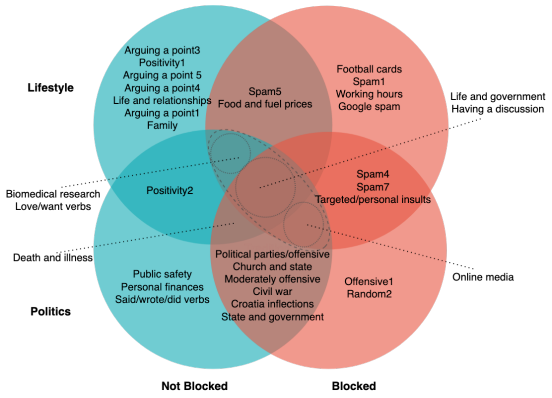


Figure 3: Top topics of the blocked and non-blocked comments of the Lifestyle and Politics sections.

and between sections by looking at the top topics of the easiest and most difficult sections, Lifestyle and Politics, respectively. Figure 3 shows the top topics of these sections and the intersections between them. In Politics, blocked comments tend toward spam and targeted insults. Non-blocked topics are about public safety, finances and scientific research. Moreover, there are many overlapping topics between blocked and non-blocked. This suggests that blocked and non-blocked comments in Politics discuss the same subjects. This supports our hypothesis that one reason why comments in Politics are difficult to classify is that thematically, blocked (excluding spam) and non-blocked comments tend to be similar. In Lifestyle, blocked topics are dominated by spam and while there are topics on offensive words and insults, they are not as prevalent as the spam ones. The non-blocked topics are about family and relationships and commenters arguing with each other. In terms of topic overlaps between Lifestyle and Politics, blocked comments in both sections are about spam and targeted insults while non-blocked comments use a more positive tone.

7.3 Analysis of Classifier Outputs

In general, we observe that blocked comments tend to use similar topics across different sections while non-blocked comments have more diverse topics. Of the 10 sections that we analyzed, there are 5 topics that are used by all blocked comments in all sections ('Targeted/personal insults', 'Spam4', 'Spam7', 'Online media', and, 'Having a discussion') and 3 topics used by all non-blocked comments ('Having a discussion', 'Online media', and, 'Life and government'). This suggests that blocked comments across sections have more in common with each other than non-blocked ones. Topics in non-blocked comments tend to be more relevant to their news section: for instance, family and relationships are not discussed a lot in the Politics section, while Lifestyle commenters do not tend to talk about the government and political parties.

In general, then, the higher topical coherence of blocked comments explains why a text classification approach can achieve reasonable performance;

Comment	Label	Text-only	LF1
konano. gamad lopovska crno bijela prevarantska (*finally. the black and white cheating thieving bastards*)	1	0 (0.034)	1 (0.687)
...dobro jutro,moze crveni karton za novinara koji je osmislio naslov ;-) (*... good morning, how about a red card for the journalist who came up with this title ;-)*)	1	0 (0.047)	0 (0.456)
Ba ste jadnici kao i ovi sa 24sata koji u ovome uivaju ! (*All of you are lame as well as those from 24sata who enjoy this.*)	1	0 (0.057)	0 (0.229)
eli pronai enu za jednu no? Dobrodoli na >>> URL (*Want to find a woman for one night? Welcome at >>> URL*)	1	1 (0.503)	1 (0.618)

Table 6: Sample comments and classifier decisions.

but the variation in blocked comment content between some sections explains why adding topic model outputs improves our classification results.

In this section, we analyze the confidence of classifiers and examine some of the outputs of the models. To analyze confidence, we gradually increase the classification threshold from 0.5 to 1.0 in increments of 0.05. For every new threshold, we plot the macro-F1 for the different models (Figure 4). We compare the confidence of four models: DTV, Text only, EF2 (the strongest early fusion model), and LF1 (the overall best-performing model). The most confident model is LF1 and the least confident is DTV. The two fusion classifiers display similar levels of confidence. The Text-only classifier is not as confident as the fusion classifiers but still more confident than DTV. This suggests that adding topic features to text not only improves performance, it also increases classifier confidence.

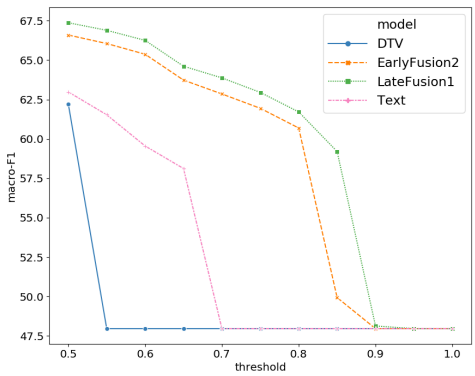


Figure 4: Confidence of the top performing models.

In Table 6 we give some examples of comments and the classifier decisions of the text-only classifier and LF1 (our best-performing fusion model). In some cases, LF1 corrects errors of the text classifier (first example); in others, the LF1 model has more confidence in the correct class (last example). The second and third examples are interesting, as LF1 has higher confidence but predicts the wrong label. However, these seem to be cases where the gold label may be incorrect (and the higher confidence justified). The second example is just a mild provocation of the moderators ("getting a red card" is an expression used for "being banned"). The third example is also only mildly offensive and not directed to anyone personally. Overall, compared to the text-only model, we find that LF1 improves the confidences (and sometimes the classification) in many cases, especially in cases which the gold label is clear. This is valuable in practice, as better confidences might lead to better prioritisation of comments for manual moderation, reducing the time required to remove the most problematic ones.

8 Conclusion

In this work, we propose a model to combine document-level semantics in the form of topics with text for comment moderation. Our analysis shows that blocked and non-blocked comments have different linguistic and thematic features, and that topics and language use vary considerably across news sections, including some variation in the comments that should be blocked. We therefore see that the use of topics in our model improves performance, and gives more confident outputs, over a model that only uses the comment text. The model also provides topic distributions, interpretable as keywords, as a form of an explanation of its prediction. As future work, we plan to incorporate comment, article, and user metadata into the model.

References

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. RuG@ EVALITA 2018: Hate Speech Detection In Italian Social Media. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:245.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Yatin Chaudhary, Hinrich Schütze, and Pankaj Gupta. 2020. Explainable and discourse topic-aware neural language understanding. In *International Conference on Machine Learning*, pages 1479–1488. PMLR.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, pages 512–515.

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.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100.

Chikashi Nobata, J. Tetreault, A. Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying toxicity within YouTube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 214–223. Springer.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017b. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55, Copenhagen, Denmark. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 166–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dietmar Schabus and Marcin Skowron. 2018. Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, pages 1241–1244.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Ravi Shekhar, Marko Pranji, Senja Pollak, Andra Pelicon, and Matthew Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Jan Šnajder. 2014. DerivBase.hr: A high-coverage derivational morphology resource for Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3371–3377, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving moderation of online discussions via interpretable neural models. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Fei Tan, Yifan Hu, Changwei Hu, Keqian Li, and Kevin Yen. 2020. TNT: Text normalization based pre-training of transformers for content moderation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4735–4741, Online. Association for Computational Linguistics.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop (GermEval)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Appendix C: Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization

Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization

Notebook for PAN at CLEF 2020

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

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

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

1 Introduction

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

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

This paper is structured as follows. In Section 2 presents related work, Section 3, we discuss the problem addressed in this work. Next, in Section 4, we discuss the proposed method, followed by empirical evaluation and discussion.

2 Related work

A critical mass of fake news can have serious, real-life consequences, and can for example impact election process [3]. Distinguishing between real and fake news content has been addressed by linguistic approaches focusing on text properties, such as the writing style and content [18] and by network approaches, where using network properties and behavior are ways to complement content-based approaches that rely on deceptive language and leakage cues to predict deception. [1] A very relevant subtopic of fake news research is detection of fake news spreaders. Commonly, fake news spreaders are implemented as bots [23], and are able to carry out the spreading process in completely *automated manner*. It is still researched, whether active prevention of fake news spreading is a viable tactic, and to what extent it can be implemented in real-life online systems [15]. Further, previous PAN submissions on the topic of bot prediction indicate (e.g., [11]), that the best models perform well when different types of textual features, entailing semantic, as well as morphological information, are used.

Twitter fake news spreaders can be captured in their own social bubbles, which was shown to be an efficient defense tactic [10]. Here, simple tweet frequency distributions were already indicative of spurious behavior. Classification via features, such as the account age and similar was also shown to work well [7]. In a recent survey [24], the authors emphasize that fact-checking is an important step in maintaining online social media *quality*. By employing automated systems, capable of prioritizing potentially interesting users, less time is spent on manual curation, which can be an expensive and time-consuming process.

Traditional classifiers with extensive feature engineering seem to be pervasive in the literature about distinguishing between bots and humans but there was also some attempts to tackle the task with neural networks. In the recent work, [5] proposed a behavior enhanced deep model (BeDM) that regards user content as temporal text data instead of plain text and fuses content information and behavior information using a deep learning method. They report an F1-score of 87.32% on a Twitter-related dataset. Finally, low-dimensional representations have recently been shown to perform well for social media-based profiling [20].

3 Problem description

Provided a timeline of chosen tweets of ground truth labeled data consisting of fake news spreaders and non-spreaders, the goal is to decide if a new author is a spreader of fake news or not. Formally, we are given a decision problem which states:
Given an author A who tweets in language $L \in \{English \vee Spanish\}$ and from the collection of tweets C, given a subset of tweets C_A (of an author A),

$$C_A = t_1, t_2, \ldots, t_n \quad \text{where} \quad t_i \quad \text{represents a tweet content,}$$

find a decision function that maps $f : C_A \mapsto$ author reliability, hence

$$f(C(A)) = \begin{cases} 0 & \text{a non fake-news spreader;} \\ 1 & \text{a fake-news spreader;} \end{cases}$$

This decision problem is specialization of the problem of *author profiling*. It requires *learning* a representation from C_A, suitable for approximating f. The provided data consists of tweets by 300 English and 300 Spanish authors respectively, respectively.

For each author 100 tweets are provided making a total of 300000 English and 300000 Spanish tweets. The balance of classes is consistent for both languages, both having 150 negative and 150 positive samples, as shown in Table 1.

Table 1. Dataset distributions

Language	spreaders	non-spreaders
English	150	150
Spanish	150	150

4 Method description

The following section includes description of the proposed method with the corresponding intermediate steps.

4.1 Pre-processing

First, the tweets from each author are concatenated, and only the printable characters are kept, which means no non-printable characters are preserved. Data pre-processing for both English and Spanish includes the following steps:

1. From the original data punctuation is removed
2. URL and hashtags are removed from the result of step (1)
3. stop-words are removed from the output of step (2).

4.2 Automatic feature construction

For each author's collection of tweets we initially define a collection of candidate n features from the pre-processed data which are iteratively selected and weighted, similarly to Martinc et. al. [12]. Features generated in the construction are based on choosing following feature types:

- character based: each of the texts is tagged with character n-grams of size 2 and 3 characters and generates a predetermined maximum allowed number of features ranging from $\frac{n}{2}$ up to 15000 features.
- word based: each of the texts is tagged with word n-grams of size 1 and 2 words and generates a preconditioned maximum allowed number of features ranging from $\frac{n}{2}$ up to 15000 features.

At this we have prepared word and character features from each author's collection of tweets, ready to be used in the feature selection step.

4.3 Dimensionality reduction via matrix factorization

Next, we perform sparse singular value decomposition (SVD)[1][8] that can be summarized via the following expression:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T.$$

The final representation (embedding) $\boldsymbol{E}$ is obtained by multiplying back only a portion of the diagonal matrix ($\boldsymbol{\Sigma}$) and $\boldsymbol{U}$, giving a low-dimensional, compact representation of the initial high dimensional matrix. Note that $\boldsymbol{E} \in \mathbb{R}^{|D| \times d}$, where d is the number of diagonal entries considered. The obtained $\boldsymbol{E}$ is suitable for a given down-stream learning task, such as classification (considered in this work). Note that performing SVD in the text mining domain is also commonly associated with the notion of *latent semantic analysis*.

4.4 Classifier selection

Classification model we aimed for in this task was to be robust yet highly flexible, one that will score well on the prepared data without using many features or extensive processing power. Following this goal we conducted a series of experiments, trying different representations with corresponding linear models as presented in Section 5. The classifiers used were the following (from scikit-learn [17]): Random Forest, Logistic Regression and the Support Vector Machines [9].

5 Conducted experiments

Considering the size of the dataset and the distribution of the data within the dataset, we preformed a series of experiments. All of them aimed to test the pipeline described in the Section 4. The experiments conducted can be divided into two main categories, based on the language considered by a given model:

1. Multilingual - Both languages' data is fused together and is subject to the same feature construction and representation creation steps.
2. Monolingual - For each language in the dataset, *English* and *Spanish*, we create a separate pipeline, that is also executed exclusively on the data from a given language.

For both approaches we performed extensive grid search over parameter space to find best hyper-parameter configuration with the help of Scikit's Learn GridSearchCV function. By doing 10-fold cross validation, the grid consisted of reducing the dimensions parametrized by k in the following interval:

$$k \in [128, 256, 512, 640, 768, 1024]$$

and the number of generated n features from the interval

$$n \in [2500, 5000, 10000, 20000, 30000].$$

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Monolingual variant was based on splitting the data from each language separately into training 90% and 10% validation set, obtaining 270 training examples $C_{training}$ and 30 validation examples $C_{validation}$. Such splits were obtained for each language. Only training data was used for feature construction and dimensionality reduction.
Multilingual variant merged the data from both languages and after that the same approach as previously was applied. Merging the data from both languages potentially reduces the computational load required to train two separate models. Data was split into training 90% and 10% validation set, obtaining 540 training examples $C_{training}$ and 60 validation examples $C_{validation}$. In each iteration we generated n features in $\mathbb{R}^{540xn}$, reduced them to dimension k obtaining a matrix from the space $\mathbb{R}^{540xk}$.

$$g(C_{training}, n \text{ features}) : \mathbb{R}^{Nxn} \xrightarrow{\text{SVD}} \mathbb{R}^{Nxk} \quad \text{where } g \text{ denotes the 4.3 process.}$$

Once constructed, the feature space was subject to learning. We experimented with both logistic regression and linear SVMs and in initially some experiments were conducted with RandomForest model, of which hyperparameters we optimized in 5-fold cross validation considering the size of the dataset. Finally, we tested the performance on the $C_{validation}$ set.

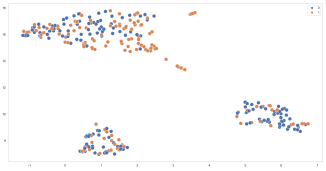


Figure 1. English Distribution

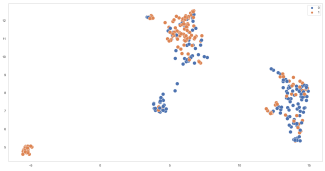


Figure 2. Spanish Distribution

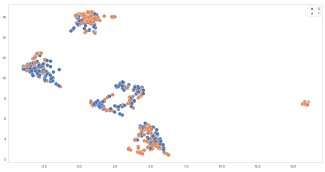


Figure 3. Merged Distribution

Figure 4. Visualization of the latent spaces used to train the final models. The orange color corresponds to spread and the blue to non-spreader. The plots indicate the number of clusters is maintained in the latent space.

We visualise the distribution of the dataset reduced to 2 dimensions using UMAP [13] dimensionality reduction in Figure 4. Figures 1 and 2 represent the visualization

with the best monolingual model described in Chapter 6, Figure 3 represents the joint latent space generated by the multilingual model described in the same chapter.

6 Results

We constructed two baselines one that was based on TF-IDF on Logistic Regression (LR) with L_1 regularization and the second was doc2vec modeled with RandomForest (RF) as classifier. The array of experiments conducted yielded the results presented in Table 2, and the outcomes of our final submission in Table 3.

As discussed in Section 5 all training was conducted by using C_{training} data and the validation was done on $C_{\text{validation}}$ set. The next presented Table 2 shows the model results as measured on TIRA training evaluation on the whole $C_{\text{validation}} \cup C_{\text{training}}$ data.

name	type	#features	#dimensions	model	EN ACC	ES ACC
tfidf_large	multi	5000	768	LR	0.9633	**0.9867**
tfidf_tweet_tokenizer	multi	5000	768	LR	0.9633	0.9533
tfidf_small	mono	5000	512	SVM,SVM	**0.9700**	0.4900
tfidf_cv	mono	10000	768	SVM,SVM	0.9100	0.9367
tfidf_no_hash	multi	10000	768	LR	0.9300	0.9067
doc2vec_baseline	mono	100	#	RF,SVM	0.6428	0.6971
tfidf_tpot_baseline	mono	30000	#	LR,SVM	0.7500	0.7400
tfidf_baseline	mono	10000	#	LR,LR	0.5567	0.7033

Table 2. Final training data on TIRA.

The final un-official evaluation as reported on TIRA's page is presented in Table 3.

name	type	#features	#dimensions	model	EN ACC	ES ACC
tfidf_large	multi	5000	768	LR	**0.7150**	**0.7950**
tfidf_cv	mono	10000	768	SVM,SVM	0.7000	**0.7950**

Table 3. Un-official evaluation on test data on TIRA

The Model column in Table 2 refers to the classifiers used, such that if two classifiers are present the model is monolingual - the first classifier is for English and the second one for Spanish and in case the model is multilingual only one classifier is used. The type column discriminates between the number of languages the model is trained on. Name column consists of vectorizer used and is followed by dimension size or type of tokinizer used or, dimensions column denotes the number of dimenstions SVD reduces to.

As it can be seen the highest evaluation score on our training data was obtained by the multilingual model *tfidf_large*, with the following hyper-parameters: $k = 768$ dimensions, $n = 5000$ features, Logistic Regression classifier with λ_2 = 0.002 and fit_intercept= $False$.

Monolingual model that preformed best is *tfidf_cv* which for English is paramatrized as SVM model with the following hyper-parameters: $\alpha = 0.001$, λ_1 = 0.8 while penalizing elastic-net, loss-function = hinge and power_t = 0.5 and for Spanish of SVM model with hyper-parameters: $\alpha = 0.0005$, λ_1 = 0.25 while penalizing elastic-net, loss-function = hinge and power_t = 0.9 .

The more detailed insight into the performance of the best performing models over the inference of the number of word and char n-grams and the accuracy on the 5fCV of the models is also given in Figures 5 and 6. The figures show the performance of the best mono and multilingual models – the confidence intervals indicate the variability obtained when repeating the experiments

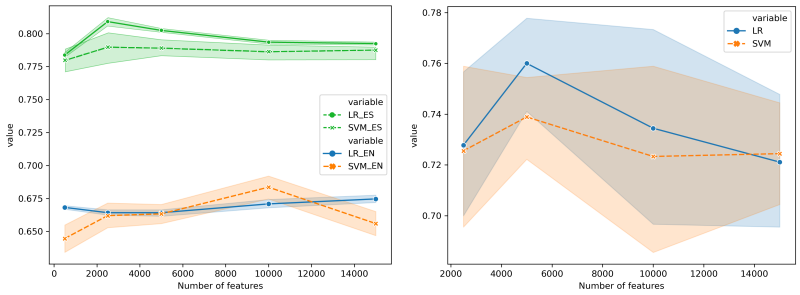


Figure 5. Best monolingual model on eval data. **Figure 6.** Best multilingual model on eval data.

7 Availability

The code and the pilot experiments are freely available at https://gitlab.com/skblaz/pan2020.

8 Discussion and Conclusions

The series of experiments conducted as a part of this work indicates, n-grams for the task of Author Profiling are still sufficient and method compared to more complex methods as transformers and word2vec [14] alike, which can easily overfit when considering only hundreds of instances. As part of the initial experiments, we also attempted to include semantic features [25], however, the results were not significantly better (nor worse), but only added to the computational time, hence such features were omitted from the final solution. We tried to change the feature space by trying different NLTK

[2] tokenizers - TweetTokenizer and the TPOT [16] automatic model generation and selection, however results obtained were similar to the ones obtained by manual construction. The joint vector space, obtained by merging the data from both languages maintains the patterns, observed when projecting individual language data sets, indicating merging of the data is a suitable tactic that does not result in complete loss of information.

Further on we can focus on exploring the possibility for detecting fake news profiles across different languages by first considering Latent Semantic Analysis across different language settings, further segmenting the semantic space prior to learning.

9 Acknowledgements

The work of the last author was funded by the Slovenian Research Agency through a young researcher grant. The work of other authors was supported by the Slovenian Research Agency (ARRS) core research programme *Knowledge Technologies* (P2-0103), an ARRS funded research project *Semantic Data Mining for Linked Open Data* (financed under the ERC Complementary Scheme, N2-0078) and European Unionś 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

1. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology Computer Science (2016)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
3. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the 2016 us presidential election. Nature communications **10**(1), 1–14 (2019)
4. Brigida, M., Pratt, W.R.: Fake news. The North American Journal of Economics and Finance **42**, 564–573 (2017)
5. Cai, C., Li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 128–130. IEEE (2017)
6. Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–18 (2020)
7. Gilani, Z., Kochmar, E., Crowcroft, J.: Classification of twitter accounts into automated agents and human users. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 489–496. ACM (2017)
8. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions (2009)
9. Hearst, M.A.: Support vector machines. IEEE Intelligent Systems **13**(4), 18–28 (Jul 1998). https://doi.org/10.1109/5254.708428, https://doi.org/10.1109/5254.708428
10. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 435–442. ACM (2010)

11. Martinc, M., Blaž Škrlj Pollak, S.: Fake or not: Distinguishing between bots, males and. CLEF 2019 Evaluation Labs and Workshop – Working Notes Papers (2019)
12. Martinc, M., Skrlj, B., Pollak, S.: Multilingual gender classification with multi-view deep learning: Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/paper_156.pdf
13. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861, https://doi.org/10.21105/joss.00861
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
15. Mustafaraj, E., Metaxas, P.T.: The fake news spreading plague: was it preventable? In: Proceedings of the 2017 ACM on web science conference. pp. 235–239 (2017)
16. Olson, R.S., Urbanowicz, R.J., Andrews, P.C., Lavender, N.A., Kidd, L.C., Moore, J.H.: Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I, chap. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137. Springer International Publishing (2016)
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
18. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3391–3401. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1287
19. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
20. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 156–169. Springer (2016)
21. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
22. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (Sep 2020), CEUR-WS.org
23. Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F.: The spread of fake news by social bots. arXiv preprint arXiv:1707.07592 **96**, 104 (2017)
24. Zhou, X., Zafarani, R.: Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315 (2018)

25. Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S.: tax2vec: Constructing interpretable features from taxonomies for short text classification. Computer Speech & Language p. 101104 (2020)

Appendix D: Identification of COVID-19 related Fake News via Neural Stacking

Identification of COVID-19 related Fake News via Neural Stacking

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

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

Please cite the original version, with the following doi: 10.1007/978-3-030-73696-5_17 (at CONSTRAINT 2021: Combating Online Hostile Posts in Regional Languages during Emergency Situation pp. 177-188) and available online at https://link.springer.com/chapter/10.1007/978-3-030-73696-5_17.

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

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

1 Introduction

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

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

2 Related Work

The fake-news text classification task [16] is defined as follows: given a text and a set of possible classes *fake* and *real*, to which a text can belong, an algorithm is asked to predict the correct class of the text. Most frequently, fake-news text classification refers to classification of data from social media. The early proposed solutions to this problem used hand crafted features of the authors such as word and character feature distributions. Interactions between fake and real news spread on social media gave the problem of fake-news detection a network-alike nature[18]. The network based modeling discovered useful components of the fake-news spreading mechanism and led to the idea of the detection of bot accounts [17].

Most of the current state-of-the-art approaches for text classification leverage large pre-trained models like the one Devlin et al. [1] and have promising results for detection of fake news [4]. However for fake-news identification tasks, approaches that make use of n-grams and the Latent Semantic Analysis [2] proved to provide successful solutions on this task (see Koloski et al. [5]). Further enrichment of text representations with taxonomies and knowledge graphs[19] promises improvements in performance.

3 Data description

In this paper we present a solution to the subset of the fake-news detection problem - The identification of COVID-19 related Fake News [11, 10]. The dataset consists of social media posts in English collected from Facebook, Twitter and Instagram, and the task is to determine for a given post if it was real or fake in relation to COVID-19. The provided dataset is split in three parts: train, validation and test data. The distribution of data in each of the data sets is shown in Table 1.

Table 1. Distribution of the labels in all of the data splits.

part	train	validation	test
size	6420	2140	2140
real	3360(52%)	1120(52%)	1120(52%)
fake	3060(48%)	1020(48%)	1020(48%)

4 Proposed method

The proposed method consists of multiple submethods that aim to tackle different aspects of the problem. On one side we focus on learning the hand crafted features of authors and on the other we focus on learning the representation of the problem space with different methods.

4.1 Hand crafted features

Word based Maximum and minimum word length in a tweet, average word length, standard deviation of the word length in tweet. Additionally we counted the number of words beginning with upper and the number of words beginning a lower case.

Char based The character based features consisted of the counts of digits, letters, spaces, punctuation, hashtags and each vowel, respectively.

4.2 Latent Semantic Analysis

Similarly to Koloski et al. [5] solution to the PAN2020-Fake News profiling we applied the low dimensional space estimation technique. First we preprocessed the data by lower-casing 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[6].

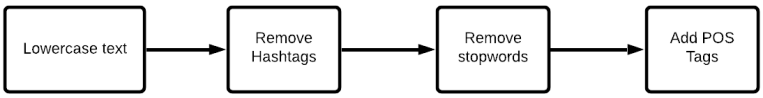


Fig. 1. Text preparation for the LSA.

For the feature construction space we used the technique used by Martinc et al. [8] which iteratievly weights and chooses the best n-grams. We used two types of n-grams:

- Word based: n-grams of size 1 and 2
- Character based: n-grams of size 1, 2 and 3

We generated n features with $n/2$ of them being word and $n/2$ character n-grams. We calculated TF-IDF on them and preformed SVD [3] With the last step we obtained the LSA representation of the tweets.

4 Koloski et al.

For choosing the optimal number of features **n** and number of dimensions **d**, we created custom grid consisted of $n` \in [500, 1250, 2500, 5000, 10000, 15000, 20000]$ and $d` \in [64, 128, 256, 512, 768]$. For each tuple $(n`, d`), n` \in \mathbf{d}$ and $d` \in \mathbf{d}$ we generated a representation and trained (SciKit library [12]) *SVM* and a *LR* (Logistic regression) classifier. The learning procedure is shown in Figure 2.

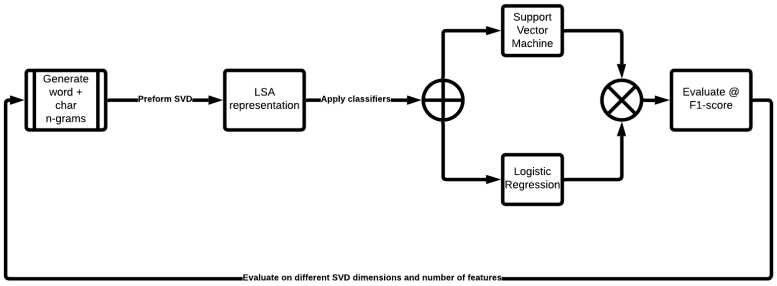


Fig. 2. The proposed learning procedure with the LSA. The evaluation is performed on validation dataset.

The best performing model had 2500 features reduced to 512 dimensions.

4.3 Contextual features

We explored two different contextual feature embedding methods that rely on the transformer architecture. The first method uses the already pretrained *sentence_transfomers* and embedds the texts in an unsupervised manner. The second method uses DistilBERT which we fine tune to our specific task.

sentence_transfomers For fast document embedding we used three different contextual embedding methods from the *sentence_transfomers* library [14]:

- *distilbert-base-nli-mean-tokens*
- *roberta-large-nli-stsb-mean-tokens*
- *xlm-r-large-en-ko-nli-ststb*

First, we applied the same preprocessing as shown in Figure 1, where we only excluded the POS tagging step. After we obtained the preprocessed texts we embedded every tweet with a given model and obtained the vector representation. After we obtained each representation, we learned a Stochastic Gradient Descent based learner, penalizing both the "linear" and "hinge" loss parameters. The parameters were optimized on a GridSearch with a 10-fold Cross-validation on every tuple of parameters.

DistilBERT is a distilled version of BERT that retains best practices for training BERT models [15]. It is trained on a concatenation of English Wikipedia and Toronto Book Corpus. To produce even better results, we fine-tuned the model on train data provided by the organizers. BERT has its own text tokenizer and is not compatible with other tokenizers so that is what we used to prepare data for training and classification.

4.4 tax2vec features

tax2vec [19] is a data enrichment approach that constructs semantic features useful for learning. It leverages background knowledge in the form of taxonomy or knowledge graph and incorporates it into textual data. We added generated semantic features using one of the two approaches described below to top 10000 word features according to the TF-IDF measure. We then trained a number of classifiers on this set of enriched features (Gradient boosting, Random forest, Logistic regression and Stochastic gradient descent) and chose the best one according to the F1-score calculated on the validation set.. **Taxonomy based (tax2vec).** Words from documents are mapped to terms of the WordNet taxonomy [13], creating a document-specific taxonomy after which a term-weighting scheme is used for feature construction. Next, a feature selection approach is used to reduce the number of features. **Knowledge Graph based (tax2vec(kg)).** Nouns in sentences are extracted with SpaCy and generalized using the Microsoft Concept Graph [9] by "is_a" concept. A feature selection approach is used to reduce the number of features.

5 Meta models

From the base models listed in Section 4 we constructed two additional meta-models by combining the previously discussed models.

5.1 Neural stacking

In this approach we learn a dense representation with 5-layer deep neural network. For the inputs we use the following representations:

- LSA representation with $N = 2500$ features reduced to $d = 256$ dimensions.
- Hand crafted features - $d = 16$ dimensions
- *distilbert-base-nli-mean-tokens* - $d = 768$ dimensions
- *roberta-large-nli-stsb-mean-tokens* - $d = 768$ dimensions
- *xlm-r-large-en-ko-nli-ststb* - $d = 768$ dimensions

This represents the final input X_{Nx2576} for the neural network. After concatenating the representations we normalized them. We constructed a custom grid consisted of learning_rate $= \lambda \in$ [0.0001, 0.005, 0.001, 0.005, 0.01, 0.05, 0.1], dropout $= p \in$ [0.1, 0.3, 0.5, 0.7], batch_size $\in$ [16,32,64,128,256], epochs $\in$

[10, 100, 1000]. In the best configuration we used the *SELU* activation function and dropout $p = 0.7$ and learning rate $\lambda = 0.001$. The loss function was *Cross-Entropy* optimized with the *StochasticGradientOptimizer*, trained on $epochs = 100$ and with $batch\_size = 32$.
Layers were composed as following:

- *input* layer - $d = 2576$ nodes
- $dense_1$ layer - $d = 896$ nodes, activation = *SELU*
- $dense_2$ layer - $d = 640$ nodes, activation = *SELU*
- $dense_3$ layer - $d = 512$ nodes, activation = *SELU*
- $dense_4$ layer - $d = 216$ nodes, activation = *SELU*
- $dense_5$ layer - $d = 2$ nodes, activation = *Sigmoid*

5.2 Linear stacking

The second approach for meta-learning considered the use of the predictions via simpler models as the input space. We tried two separate methods:

Final predictions We considered the predictions from the *LSA*, *DistilBert*, *dbert*, *xlm*, *roberta*, *tax2vec* as the input. From the models' outputs we learned a Stochastic Gradient Optimizer on 10-fold CV. The learning configuration is shown in Figure 3.

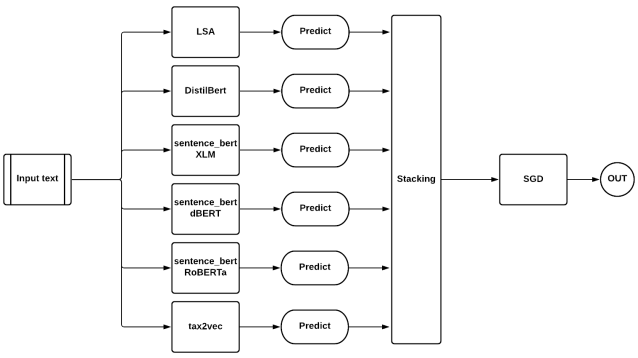


Fig. 3. Stacking architecture based on base model predictions.

Decision function-based prediction In this approach we took the given classifier's value of the decision function as the input in the stacking vector. For the SVM based SGD we used the *decision_function* and for the Logistic Regression we used the *Sidmoid_activation*. The proposed architecture is similar to the architecture in Figure 3, where prediction values are replaced by decision function values.

6 Experiments and results

This section describes model parameters, our experiments and the results of experiments as well as the results of the final submission.

We conducted the experiments in two phases. The experiment phases synced with the competition phases and were defined as *TDT* phase and *CV* phase. In the TDT phase the train and validation data is split into three subsets, while in the CV phase all data is concatenated and evaluated on 10-folds.

Table 2. Final chosen parameters for the best model of each vectorization.

Vectorization	Model	Parameters
LSA	LR	*'l1_ratio': 0.05, 'penalty': 'elasticnet', 'power_t': 0.5*
Hand crafted features	SVM	*'l1_ratio': 0.95, 'penalty': 'elasticnet', 'power_t': 0.1*
distilbert-base-nli-mean-tokens	LR	'C': 0.1, 'penalty':'l2'
roberta-large-nli-stsb-mean-tokens	LR	'C':'0.01', 'penalty': 'l2'
xlm-r-large-en-ko-nli-ststb	SVM	'C': 0.1, 'penalty': 'l2'
linear stacking_probs	SGD	'l1_ratio': 0.8, 'loss': 'hinge', 'penalty': 'elasticnet'
linear stacking	SGD	'l1_ratio': 0.3, 'loss': 'hinge', 'penalty': 'elasticnet'
tax2vec_tfidf	SGD	'alpha': 0.0001, 'l1_ratio': 0.15, 'loss': 'hinge', 'power_t': 0.5
tax2vec(kg)_tfidf	SVM	'C': 1.0, 'kernel': 'rbf'

6.1 Train-development-test (TDT) split

In the first phase, we concatenated the train and the validation data and splitted it into three subsets: *train*(75%), *dev*(18.75%) and *test*(6.25%). On the *train* split we learned the classifier which we validated on the *dev* set with measurement of F1-score. Best performing model on the *dev* set was finally evaluated on the test set. Achieved performance is presented in Table 3 and the best performances are shown in Figure 4.

Table 3. F1-scores for different methods of vectorization on the TDT data split.

Vectorization	Train F1-score	DEV F1-score	Test F1-score
distilBERT-tokenizer	**0.9933**	**0.9807**	**0.9708**
neural stacking	0.9645	0.9377	0.9461
linear stacking	0.9695	0.9445	0.9425
tax2vec	0.9895	0.9415	0.9407
linear stacking_probs	0.9710	0.9380	0.9390
LSA	0.9658	0.9302	0.9281
roberta-large-nli-stsb-mean-tokens	0.9623	0.9184	0.9142
xlm-r-large-en-ko-nli-ststb	0.9376	0.9226	0.9124
distilbert-base-nli-mean-tokens	0.9365	0.9124	0.9113
tax2vec(kg)	0.8830	0.8842	0.8892
Hand crafted features	0.7861	0.7903	0.7805

DistilBERT comes out on top in F1-score evaluation on all data sets in TDT data split—to the extent that we feared overfitting on the train data—while handcrafting features did not prove to be successful. Taxonomy based tax2vec feature construction trails distilBERTs score but using a knowledge graph to generalize constructed features seemed to decrease performance significantly (tax2vec(kg)). Other methods scored well, giving us plenty of reasonably good approaches to consider for the CV phase.

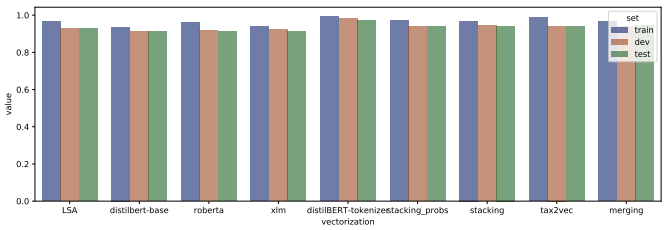


Fig. 4. Best performing methods of feature vectorization according to F1-score.

6.2 CV split

In the second phase - the *CV* phase we concatenated the data provided by the organizers and trained models on 10-fold Cross-Validation. The evaluation of the best-performing models is presented in Table 4.

During cross-validation, LSA showed consistency in good performance. With similar performance were the tax2vec methods which this time scored very similarly.

Table 4. F1-scores of models when training using 10-fold cross-validation.

Model name	Vectorization	10-fold CV
LSA	LSA	**0.9436**
sentence_transformers	distilbert	0.9071
sentence_transformers	roberta-large	0.9077
sentence_transformers	xlm-roberta	0.9123
gradient boosting	tax2vec	0.9335
gradient boosting	tax2vec(kg)	0.9350

6.3 Evaluating word features

To better understand the dataset and trained models we evaluated word features with different metrics to pinpoint features with the highest contribution to classification or highest variance.

Features with the highest variance We evaluated word features within the train dataset based on variance in *fake* and *real* classes and found the following features to have the highest variance:

"Fake" class – cure – coronavirus – video – president – covid – vaccine – trump – 19
"Real" class – number – total – new – tests – deaths – states – confirmed – cases – reported

SHAP extracted features After training the models we also used Shapley Additive Explanations [7] to extract the most important word features for classification into each class. The following are results for the gradient boosting model:

"Fake" class – video – today – year – deployment – trump – hypertext transfer protocol
"Real" class – https – covid19 – invoking – laboratories – cases – coronavirus

Generalized features We then used WordNet with a generalizing approach called ReEx (Reasoning with Explanations)[4] to generalize the terms via the "is_a" relation into the following terms:

"Fake" class – visual communication – act – matter – relation – measure – hypertext transfer protocol – attribute
"Real" class – physical entity – message – raise – psychological feature

6.4 Results

Results of the final submissions are shown in Table 5.

Table 5. Final submissions F1-score results.

submission name	model	F1-score
btb_e8_4	neural stacking	**0.9720**
btb_e8_3	LSA	0.9416
btb_e8_1	tax2vec	0.9219
btb_e8_2	linear stacking	0.8464
btb_e8_5	distilbert	0.5059

[4] https://github.com/OpaqueRelease/ReEx

DistilBERT appears to have overfitted the train data on which it achieved very high F1-scores, but failed to perform well on the test data in the final submission. Our stacking method also failed to achieve high results in the final submission, being prone to predict "fake" news as can be seen in Figure 5. On the other hand, the taxonomy based tax2vec data enrichment method as well as the LSA model have both shown good results in the final submission, while our best performing model used stacking, where we merged different neural and non-neural feature sets into a novel representation. With this merged model, we achieved 0.972 F1-score and ranked 50th out of 168 submissions.

In Figure 5 we present the confusion matrices of the models evaluated in the final submissions.

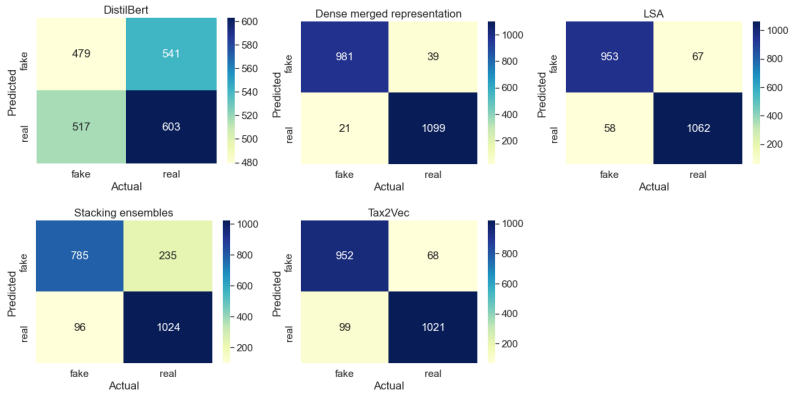


Fig. 5. Heatmaps of predicted and actual labels on final submission results.

7 Conclusion and further work

In our take to tackle the detection of fake-news problems we have have exploited different approaches and techniques. We constructed hand crafted features that captured the statistical distribution of words and characters across the tweets. From the collection of n-grams of both character and word-based features to be found in the tweets we learned a latent space representation, potentially capturing relevant patterns. With the employment of multiple BERT-based representations we captured the contextual information and the differences between fake and real COVID-19 news. However such learning showed that even though it can have excellent results for other tasks, for tasks such as classification of short texts it proved to fall behind some more sophisticated methods. To overcome such pitfalls we constructed two different meta models, learned from the decisions of simpler models. The second model learned a new space from the document space representations of the simpler models by embedding it via a 5

layer neural network. This new space resulted in a very convincing representation of this problem space achieving F1-score of **0.9720** on the final (hidden) test set.

For the further work we suggest improvements of our methods by the inclusion of background knowledge to the representations in order to gain more instance separable representations. We propose exploring the possibility of adding model interpretability with some attention based mechanism. Finally, as another add-on we would like to explore how the interactions in the networks of fake-news affect our proposed model representation.

8 Acknowledgements

The work of the last author was funded by the Slovenian Research Agency (ARRS) through a young researcher grant. The work of other authors was supported by the Slovenian Research Agency core research programme *Knowledge Technologies* (P2-0103) and the ARRS funded research projects *Semantic Data Mining for Linked Open Data (ERC Complementary Scheme, N2-0078)* and *Computer-assisted multilingual news discourse analysis with contextual embeddings - J6-2581)* . The work was also 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).

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Dumais, S.T.: Latent semantic analysis. Annual Review of Information Science and Technology **38**(1), 188–230 (2004). https://doi.org/https://doi.org/10.1002/aris.1440380105, https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440380105
3. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions (2009)
4. Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H.: exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences **9**(19), 4062 (2019)
5. Koloski, B., Pollak, S., Škrlj, B.: Multilingual detection of fake news spreaders via sparse matrix factorization. In: CLEF (2020)
6. Loper, E., Bird, S.: 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 - Volume 1. p. 63–70. ETMTNLP '02, Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1118108.1118117, https://doi.org/10.3115/1118108.1118117

7. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
8. Martinc, M., Skrlj, B., Pollak, S.: Multilingual gender classification with multi-view deep learning: Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/paper_156.pdf
9. Nsl, D.I.: Microsoft concept graph: Mining semantic concepts for short text understanding **1**, 262–294 (11 2019)
10. Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., PYKL, S., Das, A., Ekbal, A., Akhtar, M.S., Chakraborty, T.: 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)
11. Patwa, P., Sharma, S., PYKL, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Fighting an infodemic: Covid-19 fake news dataset. arXiv preprint arXiv:2011.03327 (2020)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Princeton University: About wordnet. (2010)
14. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084
15. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. CoRR **1910.01108** (2019)
16. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**(11), 2498–2504 (2003)
17. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature communications **9**(1), 1–9 (2018)
18. Shu, K., Bernard, H.R., Liu, H.: Studying fake news via network analysis: detection and mitigation. In: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, pp. 43–65. Springer (2019)
19. Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S.: tax2vec: Constructing interpretable features from taxonomies for short text classification. Computer Speech & Language p. 101104 (2020). https://doi.org/https://doi.org/10.1016/j.csl.2020.101104, http://www.sciencedirect.com/science/article/pii/S0885230820300371

Appendix E: Cross-lingual Transfer of Sentiment Classifiers

CROSS-LINGUAL TRANSFER OF SENTIMENT CLASSIFIERS

Marko ROBNIK-ŠIKONJA
Faculty of Computer and Information Science, University of Ljubljana

Kristjan REBA
Faculty of Computer and Information Science, University of Ljubljana

Igor MOZETIČ
Jožef Stefan Institute

Robnik-Šikonja, M., Reba, K., Mozetič, I. (2021): Cross-lingual transfer of sentiment classifiers. Slovenščina 2.0, 9(1): 1–25.

DOI: https://doi.org/10.4312/slo2.0.2021.1.1-25

Word embeddings represent words in a numeric space so that semantic relations between words are represented as distances and directions in the vector space. Cross-lingual word embeddings transform vector spaces of different languages so that similar words are aligned. This is done by mapping one language's vector space to the vector space of another language or by construction of a joint vector space for multiple languages. Cross-lingual embeddings can be used to transfer machine learning models between languages, thereby compensating for insufficient data in less-resourced languages. We use cross-lingual word embeddings to transfer machine learning prediction models for Twitter sentiment between 13 languages. We focus on two transfer mechanisms that recently show superior transfer performance. The first mechanism uses the trained models whose input is the joint numerical space for many languages as implemented in the LASER library. The second mechanism uses large pretrained multilingual BERT language models. Our experiments show that the transfer of models between similar languages is sensible, even with no target language data. The performance of cross-lingual models obtained with the multilingual BERT and LASER library is comparable, and the differences are language-dependent. The transfer with CroSloEngual BERT, pretrained on only three languages, is superior on these and some closely related languages.

Keywords: natural language processing, machine learning, text embeddings, sentiment analysis, BERT models

Slovenščina 2.0, 2021 (1)

1 INTRODUCTION

Word embeddings are representations of words in numerical form, as vectors of typically several hundred dimensions. The vectors are used as input to machine learning models; for complex language processing tasks, these generally are deep neural networks. The embedding vectors are obtained from specialised neural network-based embedding algorithms, e.g., fastText (Bojanowski et al., 2017) for morphologically-rich languages. Word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese (Mikolov et al., 2013). This means that embeddings independently produced from monolingual text resources can be aligned, resulting in a common cross-lingual representation, called cross-lingual embeddings, which allows for fast and effective integration of information in different languages.

There exist several approaches to cross-lingual embeddings. The first group of approaches uses monolingual embeddings with an optional help from a bilingual dictionary to align the pairs of embeddings (Artetxe et al., 2018a). The second group of approaches uses bilingually aligned (comparable or even parallel) corpora to construct joint embeddings (Artetxe and Schwenk, 2019). This approach is implemented in the LASER library[1] and is available for 93 languages. The third type of approaches is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In this work, we focus on the second and third group of approaches. In particular, from the third group, we apply two variants of BERT models, the original multilingual BERT model (mBERT), trained on 104 languages, and trilingual CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) trained on Croatian, Slovene, and English (CSE BERT).

Sentiment annotation is a costly and lengthy operation, with a relatively low inter-annotator agreement (Mozetič et al., 2016). Large annotated sentiment datasets are, therefore, rare, especially for low-resourced languages. The transfer of already trained models or datasets from other languages would increase the ability to study sentiment-related phenomena for many more languages than possible today.

1 https://github.com/facebookresearch/LASER

Our study aims to analyse the abilities of modern cross-lingual approaches for the transfer of trained models between languages. We study two cross-lingual transfer technologies, using a joint vector space computed from parallel corpora with the LASER library and multilingual BERT models. The advantage of our study is sizeable comparable classification datasets in 13 different languages, which gives credibility and general validity to our findings. Further, due to the datasets' size, we can reliably test different transfer modes: direct transfer between languages (called a zero-shot transfer) and transfer with enough fine-tuning data in the target language. In the experiments, we study two cross-lingual transfer modes based on projections of sentences into a joint vector space. The first mode transfers trained models from source to target languages. A model is trained on the source language(s) and used for classification in the target language(s). This model transfer is possible because texts in all processed languages are embedded into the common vector space. The second mode expands the training set with instances from other languages, and then all instances are mapped into the common vector space during neural network training. Besides the cross-lingual transfer, we analyse the quality of representations for the Twitter sentiment classification and compare the common vector space for several languages constructed by the LASER library, multilingual BERT models, and the traditional bag-of-words approach. The results show a relatively low decrease in predictive performance when transferring trained sentiment prediction models between similar languages and superior performance of multilingual BERT models covering only three languages.

The paper is divided into four more sections. In Section 2, we present background on different types of cross-lingual embeddings: alignment of monolingual embeddings, building a common explicit vector space for several languages, and large pretrained multilingual contextual models. We also discuss related work on Twitter sentiment analysis and cross-lingual transfer of classification models. In Section 3, we present a large collection of tweets from 13 languages used in our empirical evaluation, the implementation details of our deep neural network prediction models, and the evaluation metrics used. Section 4 contains four series of experiments. We first evaluate different representation spaces and compare the LASER common vector space with

Slovenščina 2.0, 2021 (1)

multilingual BERT models and convential bag-of-ngrams. We then analyse the transfer of trained models between languages from the same language group and from a different language group, followed by expanding datasets with instances from other languages. In Section 5, we summarise the results and present ideas for further work.

2 BACKGROUND AND RELATED WORK

Word embeddings represent each word in a language as a vector in a high dimensional vector space so that the relations between words in a language are reflected in their corresponding embeddings. Cross-lingual embeddings attempt to map words represented as vectors from one vector space to another so that the vectors representing words with the same meaning in both languages are as close as possible. Søgaard et al. (2019) present a detailed overview and classification of cross-lingual methods.

Cross-lingual approaches can be sorted into three groups, described in the following three subsections. The first group of methods uses monolingual embeddings with (an optional) help from bilingual dictionaries to align the embeddings. The second group of approaches uses bilingually aligned (comparable or even parallel) corpora for joint construction of embeddings in all handled languages. The third type of approaches is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In contrast to the first two types of approaches, the multilingual BERT models are typically used as starting models, which are fine-tuned for a particular task without explicitly extracting embedding vectors.

In Section 2.1, we first present background information on the alignment of individual monolingual embeddings. We describe the projections of many languages into a joint vector space in Section 2.2, and in Section 2.3, we present variants of multilingual BERT models. In Section 2.4, we describe related work on Twitter sentiment classification. Finally, in Section 2.5, we outline the related work on cross-lingual transfer of classification models.

2.1 Alignment of monolingual embeddings

Cross-lingual alignment methods take precomputed word embeddings for each language and align them with the optional use of bilingual dictionaries.

Two types of monolingual embedding alignment methods exist. The first type of approaches map vectors representing words in one of the languages into the vector space of the other language (and vice-versa). The second type of approaches maps embeddings from both languages into a joint vector space. The goal of both types of alignments is the same: the embeddings for words with the same meaning must be as close as possible in the final vector space. A comprehensive summary of existing approaches can be found in (Artetxe et al., 2018a). The open-source *vecmap*[2] library contains implementations of methods described in (Artetxe et al., 2018a), and can align monolingual embeddings using a supervised, semi-supervised, or unsupervised approach.

The supervised approach requires the use of a bilingual dictionary, which is used to match embeddings of equivalent words. The embeddings are aligned using the Moore-Penrose pseudo-inverse, which minimises the sum of squared Euclidean distances. The algorithm always converges but can be caught in a local maximum. Several methods (e.g., stochastic dictionary introduction or frequency-based vocabulary cut-off) are used to help the algorithm climb out of local maxima. A more detailed description of the algorithm is given in (Artetxe et al., 2018b).

The semi-supervised approach uses a small initial seeding dictionary, while the unsupervised approach is run without any bilingual information. The latter uses similarity matrices of both embeddings to build an initial dictionary. This initial dictionary is usually of low but sufficient quality for later processing. After the initial dictionary (either by seeding dictionary or using similarity matrices) is built, an iterative algorithm is applied. The algorithm first computes optimal mapping using the pseudo-inverse approach for the given initial dictionary. The optimal dictionary for the given embeddings is then computed, and the procedure iterates with the new dictionary.

When constructing mappings between embedding spaces, a bilingual dictionary can help as its entries are used as anchors for the alignment map for supervised and semi-supervised approaches. However, lately, researchers have proposed methods that do not require a bilingual dictionary but rely on the

2 https://github.com/artetxem/vecmap

Slovenščina 2.0, 2021 (1)

adversarial approach (Conneau et al., 2018) or use the words' frequencies (Artetxe et al., 2018b) to find a required transformation. These are called unsupervised approaches.

2.2 Projecting into a joint vector space

To construct a common vector space for all the processed languages, one requires a large aligned bilingual or multilingual parallel corpus. The constructed embeddings must map the same words in different languages as close as possible in the common vector space. The availability and quality of alignments in the training set corpus may present an obstacle. While Wikipedia, subtitles, and translation memories are good sources of aligned texts for large languages, less-resourced languages are not well-presented and building embeddings for such languages is a challenge.

LASER (Language-Agnostic SEntence Representations) is a Facebook research project focusing on joint sentence representation for many languages (Artetxe and Schwenk, 2019). Strictly speaking, LASER is not a word but sentence embedding method. Similarly to machine translation architectures, LASER uses an encoder-decoder architecture. The encoder is trained on a large parallel corpus, translating a sentence in any language or script to a parallel sentence in either English or Spanish (whichever exists in the parallel corpus), thereby forming a joint representation of entire sentences in many languages in a shared vector space. The project focused on scaling to many languages; currently, the encoder supports 93 different languages. Using LASER, one can train a classifier on data from just one language and use it on any language supported by LASER. A vector representation in the joint embedding space can be transformed back into a sentence using a decoder for the specific language.

2.3 Multilingual BERT and CroSloEngual BERT

BERT (Bidirectional Encoder Representations from Transformers) embedding (Devlin et al., 2019) generalises the idea of a language model (LM) to masked LMs, inspired by the cloze test, which checks understanding of a text by removing a few words, which the participant is asked to replace. The masked LM randomly masks some of the tokens from the input, and

the task is to predict the missing token based on its neighbourhood. BERT uses transformer neural networks (Vaswani et al., 2017) in a bidirectional sense and further introduces the task of predicting whether two sentences appear in a sequence. The input representation of BERT are sequences of tokens representing sub-word units. The input is constructed by summing the embeddings of corresponding tokens, segments, and positions. Some widespread words are kept as single tokens; others are split into sub-words (e.g., frequent stems, prefixes, suffixes—if needed down to single letter tokens). The original BERT project offers pre-trained English, Chinese, and multilingual model. The latter, called mBERT, is trained on 104 languages simultaneously.

To use BERT in classification tasks only requires adding connections between its last hidden layer and new neurons corresponding to the number of classes in the intended task. The fine-tuning process is applied to the whole network, and all the parameters of BERT and new class-specific weights are fine-tuned jointly to maximise the log-probability of correct labels.

Recently, a new type of multilingual BERT models emerged that reduce the number of languages in multilingual models. For example, CSE BERT (Ulčar and Robnik-Šikonja, 2020) uses Croatian, Slovene (two similar less-resourced languages from the same language family), and English. The main reasons for this choice are to represent each language better and keep sensible sub-word vocabulary, as shown by Virtanen et al. (2019). This model is built with the cross-lingual transfer of prediction models in mind. As CSE BERT includes English, we expect that it will enable a better transfer of existing prediction models from English to Croatian and Slovene.

2.4 Twitter sentiment classification

We present a brief overview of the related work on automated sentiment classification of Twitter posts. We summarise the published labelled sets used for training the classification models and the machine learning methods applied for training. Most of the related work is limited to only English texts.

To train a sentiment classifier, one needs a reasonably large training dataset of tweets already labelled with the sentiment. One can rely on a proxy, e.g.,

emoticons used in the tweets, to determine the intended sentiment; however, high-quality labelling requires the engagement of human annotators. There exist several publicly available and manually labelled Twitter datasets. They vary in the number of examples from several hundred to several thousand, but to the best of our knowledge, so far, none exceeds 20,000 entries. Saif et al. (2013) describe eight Twitter sentiment datasets and introduce a new one that contains separate sentiment labels for tweets and entities. Rosenthal et al. (2015) provide statistics for several of the 2013–2015 SemEval datasets.

There are several supervised machine learning algorithms suitable to train sentiment classifiers from sentiment labelled tweets. For example, in the SemEval-2015 competition, before the rise of deep neural networks, the most often used algorithms for the sentiment analysis on Twitter (Rosenthal et al., 2015) were support vector machines (SVM), maximum entropy, conditional random fields, and linear regression. In other cases, frequently used classifiers were naive Bayes, k-nearest neighbours, and even decision trees. Often, SVM was shown as the best performing classifier for the Twitter sentiment. However, only recently, when researchers started to apply deep learning for the Twitter sentiment classification, considerable improvements in classification performance were observed (Wehrmann et al., 2017; Jianqiang et al., 2018; Naseem et al., 2020). Similarly to our approach, recent approaches use contextual embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), but in a monolingual setting.

2.5 Transfer of trained models

Cross-lingual word embeddings can be used directly as inputs in natural language processing models. The main idea is to train a model on data from one language and then apply it to another, relying on shared cross-lingual representation. Several tasks have been attempted in testing cross-lingual transfe. Søgaard et al. (2019) survey the transfer in the following tasks: document classification, dependency parsing, POS tagging, named entity recognition, super-sense tagging, semantic parsing, discourse parsing, dialogue state tracking, entity linking (wikification), sentiment analysis, machine translation, natural language interference, etc. For example, Ranasinghe

and Zampieri (2020) apply large pretrained models in a similar way as we but use offensive language domain and only four languages from different families (English, Spanish, Bengali, and Hindu). In sentiment analysis, which is of particular interest in this work, Mogadala and Rettinger (2016) evaluate their embeddings on the multilingual Amazon product review dataset. In the Twitter sentiment analysis, Wehrmann et al. (2017) use LSTM networks but first learn a joint representation for four languages (English, German, Portuguese, and Spanish) with character-based convolutional neural networks.

3 DATASETS AND EXPERIMENTAL SETTINGS

This section presents the evaluation metrics, experimental data, and implementation details of the used neural prediction models.

3.1 Evaluation metrics

Following Mozetič et al. (2016), we report the $\overline{F}_1$ score and classification accuracy (*CA)*. The $F_1(c)$ score for class value c is the harmonic mean of precision p and recall r for the given class c, where the precision is defined as the proportion of correctly classified instances from the instances predicted to be from the class c, and the recall is the proportion of correctly classified instances actually from the class c:

$$F_1(c) = \frac{2p_c r_c}{p_c + r_c}.$$

The F_1 score returns values from the [0,1] interval, where 1 means perfect classification, and 0 indicates that either precision or recall for class c is 0. We use an instance of the F_1 score specifically designed to evaluate the 3-class sentiment models (Kiritchenko et al., 2014). $\overline{F}_1$ is defined as the average over the positive (+) and negative (–) sentiment class:

$$\overline{F}_1 = \frac{F_1(+) + F_1(-)}{2}.$$

$\overline{F}_1$ implicitly considers the ordering of sentiment values by considering only the extreme labels, positive (+) and negative (-). The middle, neutral, is taken

Slovenščina 2.0, 2021 (1)

into account indirectly. $\overline{F}_1 = 1$ implies that all negative and positive tweets were correctly classified, and as a consequence, all neutrals as well. $\overline{F}_1 = 0$ indicates that all tweets were classified as neutral, and consequently, all negative and positive tweets were incorrectly classified.

$\overline{F}_1$ is not the best performance measure. First, taking the arithmetic average of the F_1 scores over different classes (called macro F_1) is methodologically misguided (Flach and Kull, 2015). It is justified only when the class distribution is approximately even, as in our case. Second, $\overline{F}_1$ does not account for correct classifications by chance. A more appropriate measure that allows for class ordering, classification by chance, and class labelling with disagreements is Krippendorff's alpha-reliability (Krippendorff, 2013). However, since $\overline{F}_1$ is commonly used in the sentiment classification community, and the results are typically well-correlated with the alpha-reliability, we decided to report our experimental results in terms of $\overline{F}_1$.

The second score we report is the classification accuracy CA, defined as the ratio of correctly predicted tweets N_c to all the tweets N:

$$CA = \frac{N_c}{N}.$$

3.2 Datasets

We use a corpus of Twitter sentiment datasets (Mozetič et al., 2016), consisting of 15 languages, with over 1.6 million annotated tweets. The languages covered are Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovene, Spanish, and Swedish. The authors studied the annotators' agreement on the labelled tweets. They discovered that the SVM classifier achieves significantly lower score for some languages (English, Russian, Slovak) than the annotators. This hints that there might be room for improvement for these languages using a better classification model or a larger training set.

We cleaned the above datasets by removing the duplicated tweets, weblinks, and hashtags. Due to the low quality of sentiment annotations indicated by low self-agreement and low inter-annotator agreement, we removed Albanian and Spanish datasets. For these two languages, the self-agreement expressed with $\overline{F}_1$ score is 0.60 and 0.49, respectively; the inter-annotator agreement is

0.41 and 0.42. As defined above, $\overline{F}_1$ is the arithmetic average of F_1 scores for the positive and negative tweets, where $F_1(c)$ is the fraction of equally labelled tweets out of all the tweets with the label c.

In the paper where the datasets were introduced (Mozetič et al., 2016), Serbian, Croatian, and Bosnian tweets were merged into a single dataset. The three languages are very similar and difficult to distinguish in short Twitter posts. However, it turned out that this merge resulted in a poor classification performance due to a very different quality of annotations. In particular, Serbian (71,721 tweets) was annotated by 11 annotators, where two of them accounted for over 40% of the annotations. All the inter-annotator agreement measures come from the Serbian only (1,880 tweets annotated twice by different annotators, $\overline{F}_1$ is 0.51), and there are very few tweets annotated twice by the same annotator (182 tweets only, $\overline{F}_1$ for the self-agreement is 0.46). In contrast, all the Croatian and Bosnian tweets were annotated by a single annotator, and we have reliable self-agreement estimates. There are 84,001 Croatian tweets, 13,290 annotated twice, and the self-agreement $\overline{F}_1$ is 0.83. There are 38,105 Bosnian tweets, 6,519 annotated twice, and the self-agreement $\overline{F}_1$ is 0.78. The authors concluded that the annotation quality of the Croatian and Bosnian tweets is considerably higher than that of the Serbian. If one constructs separate sentiment classifiers for each language, one observes a very different performance than reported originally. The individual classifiers are better and "well-behaved" compared to the joint Serbian/Croatian/Bosnian model. In this paper, we follow the authors' suggestion that datasets with no overlapping annotations and different annotation quality are better not merged. As a consequence, the Serbian, Croatian, and Bosnian datasets are analysed separately. The characteristics of all the 13 datasets are presented in Table 1.

Slovenščina 2.0, 2021 (1)

Table 1: *The characteristics of datasets*

	Number of tweets				Agreement ($\bar{F}_1$)	
Language	**Negative**	**Neutral**	**Positive**	**All**	**Self-**	**Inter-**
Bosnian	12,868	11,526	13,711	38,105	0.78	-
Bulgarian	15,140	31,214	20,815	67,169	0.77	0.50
Croatian	21,068	19,039	43,894	84,001	0.83	-
English	26,674	46,972	29,388	103,034	0.79	0.67
German	20,617	60,061	28,452	109,130	0.73	0.42
Hungarian	10,770	22,359	35,376	68,505	0.76	-
Polish	67,083	60,486	96,005	223,574	0.84	0.67
Portuguese	58,592	53,820	44,981	157,393	0.74	-
Russian	34,252	44,044	29,477	107,773	0.82	-
Serbian	24,860	30,700	16,161	71,721	0.46	0.51
Slovak	18,716	14,917	36,792	70,425	0.77	-
Slovene	38,975	60,679	34,281	133,935	0.73	0.54
Swedish	25,319	17,857	15,371	58,547	0.76	-

Note. The left-hand side reports the number of tweets from each category and the overall number of instances for individual languages. The right-hand side contains self-agreement of annotators and inter-annotator agreement for tried languages where more than one annotator was involved.

3.3 Implementation details

In our experiments, we use three different types of prediction models, BiLSTM neural networks using joint vector space embeddings constructed with the LASER library, and two variants of BERT, mBERT, and CSE BERT. The original mBERT (bert-multi-cased) is pretrained on 104 languages, has 12 transformer layers, and 110 million parameters. The CSE BERT uses the same architecture but is pretrained only on Croatian, Slovene, and English. In the construction of sentiment classification models, we fine-tune the whole network, using the batch size of 32, 2 epochs, and Adam optimiser. We also tested larger numbers of epochs and larger batch sizes in preliminary experiments, but this did not improve the performance.

The cross-lingual embeddings from the LASER library are pretrained on 93 languages, using BiLSTM networks, and are stored as 1024 dimensional embedding vectors. Our classification models contain an embedding layer, followed by a multilayer perceptron hidden layer of size 8, and an output layer

with three neurons (corresponding to three output classes, negative, neutral, and positive sentiment) using the softmax. We use the ReLU activation function and Adam optimiser. The fine-tuning uses a batch size of 32 and 10 epochs.

Further technical details are available in the freely available source code.

4 EXPERIMENTS AND RESULTS

Our experimental work focuses on model transfer with cross-lingual embeddings. However, to first establish the suitability of different embedding spaces for Twitter sentiment classification, we start with their comparison in a monolingual setting in Section 4.1. We compare the three neural approaches presented in Section 3.3 (common vector space of LASER, mBERT, and CSE BERT). As a baseline, we use the classical approach using bag-of-ngram representation with the SVM classifier. In the cross-lingual experiments, we focus on the two most-successful types of model transfer, described in Sections 2.2 and 2.3: the common vector space of the LASER library and the variants of the multilingual BERT model (mBERT and CSE BERT). We conducted several cross-lingual transfer experiments: transfer of models between languages from the same (Section 4.2) and different language family (Section 4.3), as well as the expansion of training sets with varying amounts of data from other languages (Section 4.4). In the experiments, we did not systematically test all possible combinations of languages and language groups as this would require an excessive amount of computational time and reporting space, and would not contribute to the clarity of the paper. Instead, we arbitrarily selected a representative set of language combinations in advance. We leave a comprehensive systematic approach based on informative features (Lin et al., 2019) for further work.

4.1 Comparing embedding spaces

To establish the appropriateness of different embedding approaches for our Twitter sentiment classification task, we start with experiments in a monolingual setting. We compare embeddings into a joint vector space obtained with the LASER library with mBERT and CSE BERT. Note that there is no transfer between different languages in this experiment but only a test of

Slovenščina 2.0, 2021 (1)

the suitability of the representation, i.e. embeddings. To make the results comparable with previous work on these datasets, we report results obtained with 10-fold blocked cross-validation. There is no randomisation of training examples in the blocked cross-validation, and each fold is a block of consecutive tweets. It turns out that standard cross-validation with a random selection of examples yields unrealistic estimates of classifier performance and should not be used to evaluate classifiers in time-ordered data scenarios (Mozetič et al., 2018).

As a baseline, we report the results of SVM models without neural embeddings that use Delta TF-IDF weighted bag-of-ngrams representation with substantial preprocessing of tweets (Mozetič et al., 2016). As the datasets for the Bosnian, Croatian, and Serbian languages were merged in (Mozetič et al., 2016) due to the similarity of these languages, we report the performance on the merged dataset for the SVM classifier. Results are presented in Table 2.

Table 2: *Comparison of different representations: supervised mapping into a joint vector space with the LASER library, mBERT, CSE BERT, and bag-of-ngrams with the SVM classifier*

	LASER		mBERT		CSE BERT		SVM	
Language	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Bosnian	**0.68**	0.64	0.65	0.60	**0.68**	**0.65**	(0.61	0.56)
Bulgarian	0.53	**0.59**	**0.58**	**0.59**	0.00	0.45	0.52	0.54
Croatian	0.72	0.68	0.64	0.66	**0.76**	**0.71**	(0.61	0.56)
English	0.62	0.65	**0.68**	**0.68**	0.67	0.66	0.63	0.64
German	0.52	0.64	**0.66**	**0.66**	0.31	0.59	0.54	0.61
Hungarian	0.63	0.67	**0.65**	**0.69**	0.57	0.65	0.64	0.67
Polish	**0.70**	0.66	**0.70**	**0.70**	0.56	0.57	0.68	0.63
Portuguese	0.48	0.47	0.50	0.49	0.12	0.22	**0.55**	**0.51**
Russian	**0.70**	**0.70**	0.64	0.64	0.07	0.43	0.61	0.60
Serbian	0.50	0.54	0.50	0.52	0.30	0.50	**(0.61**	**0.56)**
Slovak	**0.72**	**0.72**	0.67	0.66	0.69	0.71	0.68	0.68
Slovene	0.57	0.58	0.58	0.58	**0.60**	**0.61**	0.55	0.54
Swedish	**0.67**	0.64	**0.67**	**0.65**	0.54	0.56	0.66	0.62
#Best	5	3	6	6	3	3	2	2

Note. The best score for each language and metric is in bold. In the last row, we count the number of best scores for each model. The SVM results for Bosnian, Croatian, and Serbian were obtained with the model trained on the merged dataset of these languages model and are therefore not directly compatible with the language-specific results for the other representations.

The SVM baseline using bag-of-ngrams representation mostly achieves lower predictive performance than the two neural embedding approaches. We speculate that the main reason is more information about the language structure contained in precomputed dense embeddings used by the neural approaches. Together with the fact that standard feature-based machine learning approaches require much more preprocessing effort, it seems that there are no good reasons why to bother with this approach in text classification; we, therefore, omit this method from further experiments. The mBERT model is the best of the tested methods, achieving the best $\overline{F}_1$ and CA scores in six languages (in bold), closely followed by the LASER approach, which achieves the best $\overline{F}_1$ score in five languages and the best CA score in three languages. The CSE BERT is specialised for only three languages, and it achieves the best scores in languages where it is trained (except in English, where it is close behind mBERT), and in Bosnian, which is similar to Croatian. Overall, it seems that large pretrained transformer models (mBERT and CSE BERT) are dominating in the Twitter sentiment prediction. The downside of these models is that their training, fine-tuning, and execution require more computational time than precomputed fixed embeddings. Nevertheless, with progress in optimisation techniques for neural network learning and advent of computationally more efficient BERT variants, e.g., (You et al., 2020), this obstacle might disappear in the future.

4.2 Transfer to the same language family

The transfer of prediction models between similar languages from the same language family is the most likely to be successful. We test several combinations of source and target languages from Slavic and Germanic language families. We report the results in Table 3.

In each experiment, we use the entire dataset(s) of the source language as the training set and the whole dataset of the target language as the testing set, i.e. we do a zero-shot transfer. We compare the results with the LASER embeddings with BiLSTM network using training and testing set from the target language, where 70% of the dataset is used for training and 30% for testing. As we use large datasets, the latter results can be taken as an upper bound of what cross-lingual transfer models could achieve in ideal conditions.

Slovenščina 2.0, 2021 (1)

The results from Table 3 (bottom line) show that there is a gap in the performance of transfer learning models and native models. On average, the gap in $\overline{F}_1$ is 5% for the LASER approach, 6% for mBERT, and 8% for CSE BERT. For CA, the average gap is 7% for both LASER and mBERT and 8% for CSE BERT. However, there are significant differences between languages, and we advise to test both LASER and mBERT for a specific new language, as the models are highly competitive. The CSE BERT is slightly less successful measured with the average performance gap over all languages as the gap is 8% in both $\overline{F}_1$ and CA. However, if we take only the three languages used in the training of CSE BERT (Croatian, Slovene, and English) as shown in

Table 3: *The transfer of trained models between languages from the same language family using LASER common vector space, mBERT, and CSE BERT*

		LASER		mBERT		CSE BERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
German	English	0.55	0.59	**0.63**	**0.64**	0.42	0.42	0.62	0.65
English	German	0.55	0.60	**0.66**	**0.70**	0.50	0.58	0.53	0.65
Polish	Russian	**0.64**	**0.59**	0.57	0.57	0.50	0.40	0.70	0.70
Polish	Slovak	**0.63**	0.59	0.58	0.59	**0.63**	**0.65**	0.72	0.72
German	Swedish	0.58	0.57	**0.59**	**0.59**	0.58	0.56	0.67	0.65
German Swedish	English	**0.58**	**0.60**	0.55	0.56	0.41	0.42	0.62	0.65
Slovene Serbian	Russian	0.53	0.55	0.57	**0.57**	**0.58**	0.48	0.70	0.70
Slovene Serbian	Slovak	**0.59**	0.52	0.57	0.59	0.48	**0.60**	0.72	0.72
Serbian	Slovene	0.54	**0.57**	0.54	0.54	**0.56**	0.55	0.60	0.60
Serbian	Croatian	**0.67**	0.64	0.65	0.62	0.65	**0.70**	0.73	0.68
Serbian	Bosnian	**0.65**	0.61	0.61	0.60	0.59	**0.62**	0.67	0.64
Polish	Slovene	0.51	0.48	**0.55**	**0.54**	0.50	0.53	0.60	0.60
Slovak	Slovene	0.52	0.51	0.54	0.54	**0.58**	**0.58**	0.60	0.60
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	Serbian	**0.54**	**0.52**	0.52	0.51	0.52	0.49	0.48	0.54
Croatian	Bosnian	0.66	0.61	0.57	0.56	**0.67**	**0.62**	0.67	0.64
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Slovene	Serbian	**0.52**	**0.55**	0.46	0.49	0.47	0.50	0.48	0.54
Slovene	Bosnian	**0.66**	0.61	0.58	0.56	**0.66**	**0.62**	0.67	0.64
Average performance gap		0.05	0.07	0.06	0.07	0.08	0.08		

Note. We compare the results with both training and testing set from the target language using the LASER approach (the right-most two columns).

Table 4, conclusions are entirely different. The average performance gap is 0% in $\overline{F}_1$ and 1% in the classification accuracy, meaning that we get almost a perfect cross-lingual transfer for these languages on the Twitter sentiment prediction task.

We also tried more than one input language at once, for example, German and Swedish as source languages and English as the target language, as shown in Table 3. The success of the tested combinations is mixed: for some models and some languages, we slightly improve the scores, while for others, we slightly decrease them. We hypothesise that our datasets for individual languages are large enough so that adding additional training data does not help.

Table 4: *The transfer of sentiment models between all combinations of languages on which CSE BERT was trained (Croatian, Slovene, and English)*

		LASER		mBERT		CSE BERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	English	**0.63**	0.63	**0.63**	**0.66**	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.50	0.53	**0.59**	**0.57**	0.60	0.60
English	Croatian	0.62	**0.67**	0.67	0.63	**0.73**	**0.67**	0.73	0.68
Slovene	English	0.63	0.64	**0.65**	**0.67**	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	**0.60**	**0.58**	0.60	0.60
Croatian Slovene	English	0.62	0.61	**0.65**	**0.67**	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	**0.68**	**0.70**	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

4.3 Transfer to a different language family

The transfer of prediction models between languages from different language families is less likely to be successful. Nevertheless, to observe the difference, we test several combinations of source and target languages from different language families (one from Slavic, the other from Germanic, and vice-versa). We compare the LASER approach with mBERT models; the CSE BERT is not constructed for this setting, and we skip it in this experiment. We report the results in Table 5.

Slovenščina 2.0, 2021 (1)

The results show that with the LASER approach, there is an average decrease of performance for transfer learning models of 11% (both $\overline{F}_1$ and CA), and for mBERT, the gap is 9%. This gap is significant and makes the resulting transferred models less useful in the target languages, though there are considerable differences between the languages.

Table 5: *The transfer of trained models between languages from different language families using LASER common vector space and mBERT*

		LASER		mBERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Russian	English	**0.52**	0.56	**0.52**	**0.57**	0.62	0.65
English	Russian	**0.57**	**0.58**	0.55	0.57	0.70	0.70
English	Slovak	0.46	0.44	**0.57**	**0.58**	0.72	0.72
Polish, Slovene	English	0.58	0.57	**0.60**	**0.60**	0.62	0.65
German, Swedish	Russian	0.61	**0.61**	**0.62**	0.59	0.70	0.70
English, German	Slovak	0.50	0.47	**0.56**	**0.54**	0.72	0.72
German	Slovene	**0.54**	**0.56**	0.53	0.54	0.60	0.60
English	Slovene	**0.54**	**0.57**	0.50	0.53	0.60	0.60
Swedish	Slovene	**0.54**	**0.56**	0.52	0.54	0.60	0.60
Hungarian	Slovene	0.52	0.52	**0.53**	**0.54**	0.60	0.60
Portuguese	Slovene	0.51	0.49	**0.54**	**0.54**	0.60	0.60
Average performance gap		0.11	0.11	0.09	0.09		

Note. We compare the results with both training and testing set from the target language using the LASER approach (the right-most two columns).

4.4 Increasing datasets with several languages

Another type of cross-lingual transfer is possible if we increase the training sets with instances from several related and unrelated languages. We conduct two sets of experiments in this scenario. In the first setting, reported in Table 6, we constructed the training set in each experiment with instances from several languages and 70% of the target language dataset. The remaining 30% of target language instances are used as the testing set. In the second setting, reported in Table 7, we merge *all* other languages and 70% of the target language into a joint training set. We compare the LASER approach, mBERT, and also CSE BERT, as Slovene and Croatian are involved in some combinations.

M. ROBNIK-ŠIKONJA, K. REBA, I. MOZETIČ: Cross-lingual transfer of sentiment classifiers

Table 6 shows a gap between learning models using the expanded datasets and models with only target language data. The decrease is more extensive for both BERT models (on average around 10%) than for the LASER approach (the decrease is on average 3% for $\overline{F}_1$ and 5% for CA). These results indicate that the tested expansion of datasets was unsuccessful, i.e. the provided amount of training instances in the target language was already sufficient for successful learning. The additional instances from other languages in the transformed space are likely to be of lower quality than the native instances and therefore decrease the performance.

Table 6: *The expansion of training sets with instances from several languages*

		LASER		mBERT		CSEBERT		Target only	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
English, Croatian, Slovene	Slovene	0.58	0.53	0.46	0.45	**0.60**	**0.58**	0.60	0.60
English, Croatian, Serbian, Slovak	Slovak	**0.67**	**0.65**	0.57	0.54	0.27	0.37	0.72	0.72
Hungarian, Slovak, English, Croatian, Russian	Russian	**0.67**	**0.65**	0.61	0.59	0.63	0.61	0.70	0.70
Russian, Swedish, English	English	0.60	0.61	**0.62**	0.60	0.59	**0.62**	0.62	0.65
Croatian, Serbian, Bosnian, Slovene	Slovene	0.54	**0.58**	0.44	0.45	**0.57**	0.56	0.60	0.60
English, Swedish, German	German	0.55	0.60	**0.60**	**0.64**	0.47	0.58	0.53	0.65
Average performance gap		0.03	0.05	0.08	0.11	0.11	0.10		

Note. We compare the LASER approach, mBERT, and CSE BERT. As the upper bound, we give results of the LASER approach trained on only the target language.

The results in Table 7, where we test the expansion of the training set (consisting of 70% of the dataset in the target language) with all other languages, show that using many languages and significant enlargement of datasets is also not successful. The two improvements in the LASER approach over using only target language are limited to a single metric (F_1 in case of Bulgarian and Serbian), which indicates that true positives are favoured at the expense of true negatives. For all the other languages, the tried expansions of training sets are unsuccessful for the LASER approach; the difference to native models

Slovenščina 2.0, 2021 (1)

is on average 3.5% for the $\overline{F}_1$ score and 6% for CA. The mBERT models are in almost all cases more successful in this massive transfer than LASER models, and they sometimes marginally beat the reference mBERT approach trained only on the target language.

Table 7: *The expansion of training sets with instances from all other languages (+70% of the target language instances) to train the LASER approach and mBERT*

	LASER				mBERT			
	All & Target		Only Target		All &Target		Only Target	
Target	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Bosnian	0.64	0.59	0.67	0.64	0.63	0.60	0.65	0.60
Bulgarian	**0.54**	0.56	0.50	0.59	**0.60**	**0.60**	0.58	0.59
Croatian	0.63	0.57	0.73	0.68	**0.65**	0.63	0.64	0.66
English	0.58	0.60	0.62	0.65	0.64	**0.69**	0.68	0.68
German	0.52	0.59	0.53	0.65	0.61	0.66	0.66	0.66
Hungarian	0.59	0.61	0.60	0.67	0.65	0.69	0.65	0.69
Polish	0.67	0.63	0.70	0.66	**0.71**	**0.71**	0.70	0.70
Portuguese	0.44	0.39	0.52	0.51	**0.52**	**0.52**	0.50	0.49
Russian	0.66	0.64	0.70	0.70	**0.67**	**0.66**	0.64	0.64
Serbian	**0.52**	0.49	0.48	0.54	**0.53**	0.51	0.50	0.52
Slovak	0.64	0.61	0.72	0.72	0.67	0.65	0.67	0.66
Slovene	0.54	0.50	0.60	0.60	0.56	0.54	0.58	0.58
Swedish	0.63	0.59	0.67	0.65	0.67	0.64	0.67	0.65
Avg. gap	0.03	0.06			0.00	0.00		

Note. We compare the results with the training on only the target language. The scores where models with the expanded training sets beat their respective reference scores are in bold.

5 CONCLUSIONS

We studied state-of-the-art approaches to the cross-lingual transfer of Twitter sentiment prediction models: mappings of words into the common vector space using the LASER library and two multilingual BERT variants (mBERT and trilingual CSE BERT). Our empirical evaluation is based on relatively large datasets of labelled tweets from 13 European languages. We first tested the success of these text representations in a monolingual setting. The results show that BERT variants are the most successful, closely followed by the LASER approach, while the classical bag-of-ngrams coupled with the SVM

classifier is no longer competitive with neural approaches. In the cross-lingual experiments, the results show that there is a significant transfer potential using the models trained on similar languages; compared to training and testing on the same language, with LASER, we get on average 5% lower $\overline{F}_1$ score and with mBERT 6% lower $\overline{F}_1$ score. The transfer of models with CSE BERT is even more successful in the three languages covered by this model, where we get no performance gap compared to the LASER approach trained and tested on the target language. Using models trained on languages from different language families produces larger differences (on average around 10% for $\overline{F}_1$ and CA). Our attempt to expand training sets with instances from different languages was unsuccessful using either additional instances from a small group of languages or instances from all other languages. The source code of our analyses is freely available[3].

We plan to expand BERT models with additional emotional and subjectivity information in future work on sentiment classification. Given the favourable results in cross-lingual transfer, we will expand the work to other relevant tasks.

Acknowledgments

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103, as well as project no. J6-2581. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153), and Rights, Equality and Citizenship Programme project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263). The results of this publication reflect only the authors' view, and the Commission is not responsible for any use that may be made of the information it contains.

3 https://github.com/kristjanreba/cross-lingual-classificationof-tweet-sentiment

REFERENCES

Artetxe, M., Labaka, G., & Agirre, E. (2018a). Generalising and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Artetxe, M., Labaka, G., & Agirre, E. (2018b). A robust self-learning method for fully unsupervised crosslingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics:Vol 1 (Long Papers)* (pp. 789–798).

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot crosslingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Conneau, A., Lample, G., Ranzato, M.A., Denoyer, L., & J'egou, H. (2018). Word' translation without parallel data. In *6th Proceedings of International Conference on Learning Representation (ICLR)*. Retrieved from https://openreview.net/pdf?id=H196sainb

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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, Vol. 1 (Long and Short Papers)* (pp. 4171–4186).

Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 838–846).

Jianqiang, Z., Xiaolin, G., and Xuejun, Z. (2018). Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access*, *6*, 23253–23260.

Kiritchenko, S., Zhu, X., Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762.

Krippendorff, K. (2013). *Content Analysis, An Introduction to Its Methodology* (3rd ed.) Thousand Oaks, CA, USA: Sage Publications.

Lin, Y. H., Chen, C. Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., et al. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 3125–3135).

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint* 1309.4168.

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL-HLT* (pp. 692–702).

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE, 11*(5). doi: 10.1371/journal.pone.0155036

Mozetič, I., Torgo, L., Cerqueira, V., & Smailović, J. (2018). How to evaluate sentiment classifiers for Twitter time-ordered data? *PLoS ONE 13*(3).

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for Twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58–69.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualised word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers)* (pp. 2227–2237).

Ranasinghe, T., & Zampieri, M. (2020). Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5838–5844).

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., & Stoyanov, V. (2015). SemEval-2015 task 10: Sentiment Analysis in Twitter. In *Proceedings of 9th International Workshop on Semantic Evaluation (SemEval)* (pp. 451–463).

Saif, H., Fernández, M., He, Y., Alani, H.(2013). Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In *1st Intl. Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM)*.

Slovenščina 2.0, 2021 (1)

Søgaard, A., Vulić, I., Ruder, S., & Faruqui, M. (2019). *Cross-Lingual Word Embeddings*. Morgan & Claypool Publishers.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue (TSD)* (pp. 104–111).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 5998–6008).

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luoto-lahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint* 1912.07076.

Wehrmann, J., Becker, W., Cagnini, H. E., & Barros, R. C. (2017). A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. In 2017 *International Joint Conference on Neural Networks (IJCNN)* (pp. 2384–2391).

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., et al. (2020). Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations (ICLR), 26-30 April, 2020, Addis Ababa, Ethiopia*.

M. ROBNIK-ŠIKONJA, K. REBA, I. MOZETIČ: Cross-lingual transfer of sentiment classifiers

MEDJEZIKOVNI PRENOS KLASIFIKATORJEV SENTIMENTA

Vektorske vložitve predstavijo besede v številski obliki tako, da so semantične relacije med besedami zapisane kot razdalje in smeri v vektorskem prostoru. Medjezikovne vložitve poravnajo vektorske prostore različnih jezikov, kar podobne besede v različnih jezikih postavi blizu skupaj. Medjezikovna poravnava lahko deluje na parih jezikov ali s konstrukcijo skupnega vektorskega prostora več jezikov. Medjezikovne vektorske vložitve lahko uporabimo za prenos modelov strojnega učenja med jeziki in s tem razrešimo težavo premajhnih ali neobstoječih učnih množic v jezikih z manj viri. V delu uporabljamo medjezikovne vložitve za prenos napovednih modelov strojnega učenja za napovedovanje sentimenta tvitov med trinajstimi jeziki. Osredotočeni smo na dva, v zadnjem času najuspešnejša, načina prenosa modelov. Prvi način uporablja modele naučene na skupnem vektorskem prostoru za mnoge jezike, izdelanem s knjižnico LASER. Drugi način uporablja velike, na mnogih jezikih vnaprej naučene, jezikovne modele tipa BERT. Naši poskusi kažejo, da je prenos modelov med podobnimi jeziki smiseln tudi povsem brez učnih podatkov v ciljnem jeziku. Uspešnost večjezikovnih modelov BERT in LASER je primerljiva, razlike so odvisne od jezika. Medjezikovni prenos z modelom CroSloEngual BERT, predhodno naučenim na le treh jezikih, je v teh in nekaterih sorodnih jezikih še precej boljši.

Ključne besede: obdelava naravnega jezika, strojno učenje, vektorske vložitve besedil, analiza sentimenta, modeli BERT

Appendix F: Temporal Mental Health Dynamics on Social Media

Temporal Mental Health Dynamics on Social Media

Tom Tabak[1] **Matthew Purver** [1 2]
[1]School of Electronic Engineering and Computer Science
Queen Mary University of London
[2]Department of Knowledge Technologies, Jožef Stefan Institute
tabaktom360@gmail.com m.purver@qmul.ac.uk

Abstract

We describe a set of experiments for building a temporal mental health dynamics system. We utilise a pre-existing methodology for distant-supervision of mental health data mining from social media platforms and deploy the system during the global COVID-19 pandemic as a case study. Despite the challenging nature of the task, we produce encouraging results, both explicit to the global pandemic and implicit to a global phenomenon, Christmas Depression, supported by the literature. We propose a methodology for providing insight into temporal mental health dynamics to be utilised for strategic decision-making.

1 Introduction

Mental health issues pose a significant threat to the general population. Quantifiable data sources pertaining to mental health are scarce in comparison to physical health data (Coppersmith et al. 2014). This scarcity contributes to the complexity of development of reliable diagnoses and effective treatment of mental health issues as is the norm in physical health (Righetti-Veltema et al. 1998). The scarcity is partially due to complexity and variation in underlying causes of mental illness. Furthermore, the traditional method for gathering population-level mental health data, behavioral surveys, is costly and often delayed (De Choudhury, Counts & Horvitz 2013*b*).

Whilst widespread adoption and engagement in social media platforms has provided researchers with a plentiful data source for a variety of tasks, including mental health diagnosis; it has not, yet, yielded a concrete solution to mental health diagnosis (Ayers et al. 2014). Conducting mental health diagnosis tasks on social media data presents its own set of challenges: The users' option of conveying a particular public persona posts that may not be genuine; sampling from a sub-population that is either technologically savvy, which may lend to a generational bias, or those that can afford the financial cost of the technology, which may lead to a demographic bias. However, the richness and diversity of the available data's content make it an attractive data source. Quantifiable data from social media platforms is by nature social and crucially (in the context of our cases study) virtual.

Quantifiable social media data enables researchers to develop methodologies for distant mental health diagnosis and analyse different mental illnesses (De Choudhury, Counts & Horvitz 2013*a*). Distant detection and analysis enables researchers to monitor relationships of temporal mental health dynamics to adverse conditions such as war, economic crisis or a pandemic such as the Coronavirus (COVID-19) pandemic.

COVID-19, a novel virus, proved to be fatal in many cases during the global pandemic that started in 2019. Governments reacted to the pandemic by placing measures restricting the movement of people on and within their borders in an attempt to slow the spread of the virus. The restrictions came in the form of many consecutive temporary policies that varied across countries in their execution. We focus on arguably the most disruptive measure: The National Lockdown. This required individuals, other than essential workers (e.g. healthcare professionals) to remain in their own homes. The lockdown enforcement varied across countries but the premise was that individuals were only permitted to leave their homes briefly for essential shopping (food and medicine). This policy had far reaching social and economic impacts: growing concern towards individuals' own and their families' health, economic well-being and financial uncertainty as certain industries (such as hospitality, retail and travel) suspended operations. As a result, many individuals became redundant and unemployed which constrained their financial re-

sources as well as being confined to their homes, resulted in excess leisure time. These experiences along with the uncertainty of the measures' duration reflected a unique period where the general public would be experiencing a similar *stressful* and *anxious* period, which are both feelings associated with clinical depression (Rickels & Schweizer 1993, Hecht et al. 1989).

In this paper, we investigate the task of detecting whether a user is diagnosis-worthy over a given period of time and explore what might this appropriate time period be. We investigate the role of balance of classes in datsets by experimenting with a variety of training regimes. Finally, we examine the temporal mental health dynamics in relations to the respective national lockdowns and investigate how these temporal mental health dynamics varied across countries highly-disrupted by the pandemic. Our main contributions in this paper are: 1) We demonstrate an improvement in mental health detection performance with increasingly enriched sample representations. 2) We highlight the importance of the balance in classes of the training dataset whilst remaining aware of an approximated expected balance of classes in the unsupervised (test) dataset. 3) We analyse empirically proven relationships between populations' temporal mental health dynamics and respective national lockdowns that can be used for strategic decision-making purposes.

2 Related research

2.1 Natural Language Processing for Mental Health Detection

Unlike physical health conditions that often show physical symptoms, mental health is often reflected by more subtle symptoms (De Choudhury, Counts & Horvitz 2013*a*, Chung & Pennebaker 2007). This yielded a body of work that focused on linguistic analysis of lexical and semantic uses in speech, such as diagnosing a patient with depression and paranoia (Oxman et al. 1982). Furthermore, an examination of college students' essays, found an increased use of negative emotional lexical content in the group of students that had high scores on depression scales (Rude et al. 2004). Such findings confirmed that language can be an indicator of an individual's psychological state (Bucci & Freedman 1981) which lead to the development of Linguistic Enquiry and Word Count (LIWC) software (Pennebaker et al. 2003, Tausczik & Pennebaker 2010) which allows users to evaluate texts based on word counts in a variety of categories. More recent and larger scale computational linguistics have been applied in conversational counselling by utilising data from an SMS service where vulnerable users can engage in therapeutic discussion with counsellors (Althoff et al. 2016). For a more in-depth review of uses of natural language processing (NLP) techniques applied in mental health the reader is referred to Trotzek et al. (2018).

2.2 Social Media as a Platform for Mental Health Monitoring

The widespread engagement in social media platforms by users coupled with the availability of platforms' data enables researchers to extract population-level health information that make it possible to track diseases, medications and symptoms (Paul & Dredze 2011). The use of social media data is attractive to researchers not only due to its vast domain coverage but also due to the cheap methodologies by which data can be collected in comparison to previously available methodologies (Coppersmith et al. 2014). A plethora of mental health monitoring literature have utilised this cheap and efficient data mining methodologies from a variety of social media platforms such as: Reddit (Losada & Crestani 2016), Facebook (Guntuku et al. 2017) and Twitter (De Choudhury, Gamon, Counts, & Horvitz 2013).

Twitter user's engagement in the popular social media platform give way for the creation of social patterns that can be analysed by researchers, making this platform a widely used data source for data mining. Additionally, the customisable parameters querying available in the Application Programmable Interface (API) allows researchers to monitor specific populations and/or domains (De Choudhury, Counts & Horvitz 2013*b*).

2.3 Mental Health Monitoring During COVID-19 Pandemic

In the context of the COVID-19 pandemic, we found a handful of projects with similar intentions as our own, to monitor depression during the pandemic. Li et al. (2020) gather large scale, pandemic-related twitter data and infers depression based on emotional characteristics and sentiment analysis of tweets. Zhou et al. (2020) focus on detecting community level depression in Australia during the pandemic. They use the distant-supervision methodologies of Shen et al. (2017) to gather a

balanced dataset, they utilise the methodology of Coppersmith et al. (2014) to model the rates of depression and observing the relationship with the number of COVID-19 infections in the community. Our work differs from this in three main areas: 1) We investigate the implication of different sample representations to provide more context to our classifier. 2) We retain an imbalance in our development dataset. 3) We investigate European countries (France, Germany, Italy, Spain and the United Kingdom) that experienced a relatively high number of COVID-19 infections.

3 Diagnosis Classifier Experiments

We describe the data mining methodology used to build a distantly supervised dataset and the classifier experiments conducted on this dataset.

3.1 Data

To conduct the proposed experiments, we construct a distantly supervised development dataset for each country, to be used in training and validation of the classifier. The data mining methods follow the novel distant-supervision methodology proposed in Coppersmith et al. (2014) as it is relatively cheap but also well-structured for clinical experiments.

We follow the wide-accepted methodology proposed by Watson (1768) where diagnosed (`Diagnosed`) and non-diagnosed (`Control`), groups are created. In this paper we will only be exploring depression as a mental health condition, accordingly we will have a single `Diagnosed` group for each country's development dataset. However, if multiple mental issues were to be explored, then the same number of different `Diagnosed` groups would be required for each country's dataset.

3.1.1 `Diagnosed` Group

We gather 200 public tweets with a geolocation inside the country of interest, posted during a two-week period (1 July 2019 - 15 July 2019). As we are searching for a *depression* `Diagnosed` tweets, this two-week period needs to be chosen strategically, as we want to capture users that have been diagnosed with depression rather than seasonal affect disorder (SAD), a separate albeit a condition with similar symptoms. Tweets collected via Twitter's API[1], were retrieved based on lexical content indicating that the user has history/is currently dealing with a clinical case, e.g. "I was diagnosed with depression", rather than expressing depression in a colloquial context. Human annotators were then instructed to remove tweets that are perceived to have made a non-genuine statement regarding the users' own diagnosis, most of these were referring to a third party. Examples of genuine and non-genuine tweets encountered can be seen in *Table 1*.

Table 1: Annotation Example

Diagnosis indication	Example tweet
Genuine	"I was diagnosed with severe depression and went through the works of treatment for it."
Non-genuine	"It's official. My guinea pig has been diagnosed with depression"

We then collect all (up to 5,000 most recent) tweets made public by the remaining users between the start of 2015 and October 2019. Further filtering includes removal of all users with less than 20 tweets during this period or those whose tweets do not meet our major language of instruction benchmark. This benchmark requires 70% of the tweets collected to be written in the major language of instruction of the country of interest (i.e. United Kingdom is English, Italy is Italian etc.). Following this filtering process and some preprocessing on the tweet level, which includes medial capital splitting, mention white-space removal (i.e. if another user was mentioned this will be shown as a unique *mention* token), the same has been done with URLs, all uppercase and non-emoticon related punctuation were removed.

3.1.2 `Control` Group

We gather 10,000 public tweets with a geolocation in the country of interest, posted during the same two-week period as `Diagnosed` in 2019 and remove any tweets made by `Diagnosed` users. We then follow a similar process to that of `Diagnosed` collection methodology by collecting up to 5,000 most recent tweets for each user from the period mentioned above.

As can be seen in *Table 2*, we construct imbalanced datasets. World Health Organisation (WHO) claim 264 million people suffer from depression worldwide[2]. Whilst, at the time of writing, the

[1]Twitter API: https://developer.twitter.com/en/docs

[2]World Health Organization, "Depression," 2020, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression[Accessed: 26 July 2020]

Table 2: Composition of Development Datasets

Country	Group	No. Users	No. Tweets
France	Diagnosed	57	190,447
	Control	1,041	2,861,580
Germany	Diagnosed	53	160,864
	Control	1,138	2,802,959
Italy	Diagnosed	38	132,743
	Control	1,051	2,514,483
Spain	Diagnosed	53	107,833
	Control	1,013	2,564,966
U.K.	Diagnosed	98	289,624
	Control	1,365	3,319,201

global population stands at approximately 7.8 billion[3]. This would suggest that 1 in 30 individuals suffer from depression. However, these figures are approximations. Therefore, the extent to which our datasets are imbalanced is not an attempt to create datasets that are representative of the expected balance of classes, as these are unverifiable. Nevertheless, our datasets present ratios of Control:Diagnosed samples between 23.78:1 and 11.46:1, which came about from the data mining methods previously described. We accept these ratios to retain imbalanced datasets in a similar order of magnitude as the expected balance whilst achieving reasonable classifier performance.

3.1.3 Caveats

We inherit the limitations of the distant-supervision approach of Coppersmith et al. (2014):

1. When sampling a population we always run the risk of only capturing a subpopulation of Control or Diagnosed that is not fully representative of the population, especially considering that the Diagnosed group are identified based a single affirming tweet about an intimate subject – this attribute may not generalise well to the entire population.
2. We supervise all tweets of a unique user based on a single affirming tweet. Hence, this may result in different tweets with identical, or similar, meaning representations being assigned different labels.
3. We do not implement a verification of the method used to identify users in the Diagnosed group but rather rely on the social stigma around mental illness whereas it could be regarded as unusual for a user to tweet about a diagnosis of a mental health illness that is fictitious.
4. Control is likely contaminated with users that are diagnosed with a variety of conditions, perhaps mental health related, whether they explicitly mention this or not. We have made no attempt to remove such users.
5. Twitter users may not be entirely representative sample of the population.

3.2 Methodology

We describe the experiments conducted in classifier training of depression diagnosis. The trained classifier is deployed in *Section 4* for classifying samples from an unsupervised experiment dataset which is then used in analysing temporal mental health dynamics.

3.2.1 Sample Representation

We investigate the most appropriate sample representation of our distantly supervised dataset. We are posed with these considerations:

1. Symptoms' temporal dependencies: as the tweets gathered come from a variety of days, weeks, months and even years, symptoms may only be present in specific time-dependant samples. However, when represented by overwhelming tweet-enriched samples the classifier performance is traded-off with retaining the symptoms' temporal dependencies.
2. As our final task will be to monitor and analyse the temporal mental health dynamics, we are interested in modelling the rate of depression as fine-grained as possible.

Therefore, the ability to accurately identify Diagnosed samples and correctly discriminate between Control and Diagnosed with the least tweet-enriched samples will be vital in modelling a fine-grained rate of depression in the deployment stage of the final task where conclusions could be drawn in the context of the national lockdowns. The sample representations we examined:

- $Individual$ – each sample constitutes of a single tweet.
- $User\ day$ - each sample constitutes of all tweets by a unique user during a given day.

[3] Worldometer. 2020. Worldometer - Real Time World Statistics. [online] Available at: https://www.worldometers.info [Accessed 19 August 2020].

- $User\ week$ – each sample constitutes of all tweets by a unique user during a given week.
- $All\ user$ - each sample constitutes of all tweets collected from a unique user.

We examine the performance of a benchmark, Support Vector Machine (SVM) with a linear kernel function (Peng et al. 2019), on the different sample representations datasets where the benchmark classifier inputs are sparse many-hot encoding representations of the samples' lexical content. As we are working with imbalanced datasets we need to think about the metrics we use to assess the classifiers' performance. We will be assessing class specific Precision (P) and Recall (R) as well as Macro F1 score. By having a more class-specific breakdown of the classifiers' performance we can better understand the strengths and limitations of our classifiers and hence make a more informed decision when choosing the highest performing classifier.

Table 3: SVM Performance on Varying Sample Representations of U.K. Development Datasets

Sample Representation	Control			Diagnosed			Macro
	P	R	F1	P	R	F1	F1
*Individual**	0.92	0.99	0.95	0.27	0.05	0.08	0.52
User day	0.74	0.96	0.84	0.36	0.06	0.1	0.52
User week	0.92	0.97	0.94	0.26	0.11	0.15	0.55
All user	0.94	0.99	0.96	0.5	0.14	0.22	0.59

The results in *Table 3* suggest that our benchmark classifier improved in identifying `Diagnosed`, with increasingly tweet-enriched, samples. However, the $User\ day$ sample representations shows a decrease in performance when compared with the F1 scores of both $Individual$ and $User\ week$ sample representations. Barring this decrease, we can say that we are able to achieve improved performance when using increasingly tweet-enriched samples. However, our final task is bias towards the two fine-grained sample representations, $Individual$ and $User\ day$. The benchmark classifier achieves superior performance on the $Individual$ sample representation, we will adopt this representation in further experiments, as denoted by the asterisk in *Table 3*.

3.2.2 Classifier Experiments on U.K. Development Dataset

We must now build and train a classifier architecture that best discriminates between our two classes. Classifier architectures included in our experimentation: SVM: SVM as used in *Section 3.2.1*. This classifier will serve as our benchmark; $AVEPL_{EFC}$ [4]: Average pooling layer; $CNN\text{-}MXPL_{EFC}$: CNN[5] and a Max-pooling layer; $BILSTM_{EFC}$: Bi-directional LSTM (Hochreiter & Schmidhuber 1997); $CNN\text{-}BILSTM_{EFC}$: CNN and Bi-directional LSTM; $CNN\text{-}ATT$: CNN, Attention (Vaswani et al. 2017) and Average pooling layers; $BILSTM\text{-}SELFA$: Bi-directional LSTM and Self-attention layer; and $BERT$: Pretrained $BERT_{Base}$[6] fine-tuned on our dataset (Devlin et al. 2018). All classifiers use an Adam optimiser (Kingma & Ba 2015) and were trained for a single epoch on a training:validation split of 4:1 with weighting the `Diagnosed` samples as 5 times more valuable than those of `Control`. Training on a single epoch was chosen in line with our theme of quick development (also conveyed in the distant-supervision), we argue that performance could be improved by further training. The sample weighting factor was chosen following empirical evidence showing that the chosen factor yielded similar `Diagnosed` Precision and Recall measures.

Table 4: Classifiers' Performance on U.K. Development Dataset

Classifier	Control			Diagnosed			Macro
	P	R	F1	P	R	F1	F1
SVM	0.92	0.9	0.91	0.27	0.05	0.08	0.52
AVEPL	0.94	0.95	0.94	0.33	0.26	0.29	0.62
CNN-MXPL	0.94	0.95	0.94	0.31	0.25	0.28	0.61
BILSTM	0.94	0.95	0.94	0.3	0.33	0.31	0.63
CNN-BILSTM	0.93	0.97	0.95	0.3	0.15	0.2	0.57
CNN-ATT	0.94	0.94	0.94	0.31	0.3	0.3	0.62
*BILSTM-SELFA**	0.94	0.94	0.94	0.32	0.31	0.31	0.63
BERT	0.94	0.66	0.78	0.12	0.52	0.2	0.47

Table 4 shows all classifiers achieve significantly higher performance on `Control` than the `Diagnosed`. The first observation is with respect to the poor performance of $BERT$ with a *Macro F1* of 0.47, lower than our benchmark SVM classifier. We argue that this poor performance can be attributed to the extensively trained word embeddings of the $BERT$ classifier remaining under-

[4] All uses of X_{EFC} indicate a learned embedding layer, after the input, and 3 Fully Connected layers, with Rectified Linear (ReLU), directly prior to the output layer.

[5] All CNNs are unigram-level with 1 $filter$ and $kernel\ size$ of 1.

[6] Exact pretrained $BERT_{Base}$ version implementation available here: https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

utilised due to our classifier's input, which contains many spelling errors. Admittedly, this hypothesis is mere conjecture and we leave this topic for future work. As we are trying to correctly detect `Diagnosed` samples and discriminate between the two classes, we prioritise the `Diagnosed` ***Precision*** and ***Macro F1*** score metrics. Based on these 2 chosen metrics to guide our classifier selection process 3 candidates emerge: $AVEPL$, $BILSTM$ and $BILSTM\text{-}SELFA$ achieving {`Diagnosed` ***Precision***, ***Macro F1***} scores of: {0.33, 0.62}; {0.3, 0.63} and {0.32, 0.63} respectively. Whilst the performance of these classifiers is similar, $BILSTM\text{-}SELFA$ is the highest performance combination of the desired metrics (indicated by the asterisk) and as such we will be adopting this classifier in further experiments.

3.2.3 Dataset Balance Experiment

In this section we investigate the distribution of our datasets in training and validation of our classifier. By conducting this experiment we intend to gather an in-depth understanding of our task from a linguistic standpoint. We train and validate the classifier on datasets with varying balances to investigate the role of our imbalanced dataset in the depression diagnosis task. This experiment analyses the performance of the $BILSTM\text{-}SELFA$ classifier on a number of different training regimes:

- $Balanced$: a dataset containing all `Diagnosed` samples and downsampling from `Control`.
- $Imbalanced$: a dataset of the development dataset's distribution (See *Table 2*).

Furthermore, we explore the effects of sample weighting by weighting `Diagnosed` samples as 5 times more valuable than the `Control` samples as mentioned in *Section 3.2.2*. The performance of the $BILSTM\text{-}SELFA$ classifier on the different training regimes can be seen in *Table 5*.

Table 5: $BILSTM\text{-}SELFA$ Performance on Varying Training Regimes

Training	Validation	Sample Weighting	`Control`			`Diagnosed`			Macro
			P	R	F1	P	R	F1	F1
Balanced	*Balanced*	None	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Imbalanced	*Imbalanced*	None	0.93	1	0.96	0.72	0.11	0.19	0.58
Imbalanced	*Imbalanced*	Weighted	0.94	0.94	0.94	0.32	0.31	0.31	0.63
Balanced	*Imbalanced*	None	0.95	0.63	0.76	0.14	0.66	0.23	0.49
Balanced	*Imbalanced*	Weighted	0.99	0.13	0.23	0.09	0.98	0.16	0.2
Imbalanced	*Balanced*	None	0.53	1	0.69	0.98	0.11	0.2	0.45
Imbalanced	*Balanced*	Weighted	0.58	0.96	0.72	0.88	0.29	0.44	0.58

The $Balanced\text{-}Balanced$ training regime achieves an encouraging Precision-Recall trade-off, for both classes, as well as the Macro F1 score. This shows that the problem is reasonably linguistically achievable, when the imbalance challenge is removed. The $Imbalanced\text{-}Imbalanced$ training regime shows that adjusting the sample weighting is a successful measure that we can implement to adjust the Precision-Recall trade-off in our class of interest (`Diagnosed`). Our classifier performs significantly worse in the $Balanced\text{-}Imbalanced$ regime when compared to the performance on the $Imbalanced\text{-}Imbalanced$ regime, this performance is reduced by the introduction of sample weighting. Therefore, when training on a $Balanced$ dataset our classifier is less robust to an $Imbalanced$ validation dataset. Finally, whilst our classifier experiences a significant improvement in performance on the $Imbalanced\text{-}Balanced$ training regime when sample weighting is introduced due to our final depression diagnosis task in which we expect an $Imbalanced$ dataset (see *Section 3.1.2*) the training regimes implementing $Balanced$ validation datasets are not suitable approximations of our classifier's depression diagnosis performance. Therefore, the $Imbalanced$ training, with suitable sample weighting, yields more desirable and robust depression diagnosis performance as it is exposed to a broader range of data examples in training (i.e. no sub-sampling).

3.3 Results

We train separate $BILSTM\text{-}SELFA$ classifiers on each of the respective countries' imbalanced development datasets following the $Individual$ sample representation (see *Table 7 Appendix A.1*). We observe that the $BILSTM\text{-}SELFA$ architecture achieved similar performance on the remaining countries' datasets. Whilst the $BILSTM\text{-}SELFA$ classifier architecture achieved the highest performance of all our classifier architectures, a combination of 0.32 `Diagnosed` ***Precision*** and 0.63 ***Macro F1*** leaves much to be desired. As such, we perform an error analysis and examine the significance of the results.

3.3.1 Error Analysis

Table 6 shows the input samples, *Text*, the *Prediction type* as well as the *Sigmoid Output* which is the output layer of the classifier and is responsible for the final classification of the samples. The *Sigmoid Output* is normalised in the range of

$[0, 1] \in \mathbb{R}$, where an output of 0.5 represent the decision boundary, and is interpreted as complete uncertainty with regards to the sample's classification. A *Sigmoid Output* of 1 is complete certainty that the sample should be classified as `Diagnosed` and 0 is complete certainty in `Control`.

Table 6: Classification Examples for Error Analysis

Prediction Type	Text	Sigmoid Output
True Positive	"hi davenport handmade is a small one man business i make handmade wooden bowls pens jewellery boxes and other wooden items in a workshop that i built myself it started as a way of overcoming depression and has taken over my life"	0.999
False Positive	"im too depressed lol"	0.507
False Negative	"i miss you too man its actually depressing me"	0.19
True Negative	"half term kids camps are up on wandsworth common with a dedicated kids football camp"	0.001

The true positive example mentions having "overcoming depression" which implies that the user has recovered from depression, as one overcomes other health issues. The *Sigmoid Output* is 0.999 which is extremely high certainty by the classifier that this is a `Diagnosed`. Whilst, the true negative is unrelated to depression nor its underlying symptoms, as such it is classified as part of `Control` with a *Sigmoid Output* of 0.001. However, the *Texts* of the misclassified samples are similar. Both use words stemming from the word 'depress' in colloquial contexts, with no indication of clinical appropriations of depression. The *Sigmoid Outputs* of these samples are less polarised than those correctly classified, the *Sigmoid Output* of the false positive sample is marginally misclassified. However, these misclassified samples reflect the complexity of the task.

3.3.2 Significance of Results

We perform a χ^2 significance test to investigate the significance of our classifiers' results. Our null hypothesis, H_0, states that both sets of data, our classifiers' predictions ($\mathcal{D}_P$) and the distribution it is being tested against ($\mathcal{D}_T$), have been drawn from the same distribution ($\mathcal{D}$).

$$H_0 : \mathcal{D}_P \cap \mathcal{D}_T \subseteq \mathcal{D} \tag{1}$$

We compare the distribution of the classifiers' predictions against a random uniformly distributed set (Uniform) and against a random distributed set following the distribution of the development datasets (Weighted). All classifier results in *Table 7* are statistically significant from the random baselines, according to the χ^2 significance test (see *Table 8* in *Appendix A.2*). Therefore, we reject H_0 and conclude that the classifiers' predictions and those of the respective randomly distributed benchmarks have not been drawn from the same distribution.

4 Monitoring and Analysis

We prepare the unsupervised dataset and deploy the previously trained $BILSTM\text{-}SELFA$ classifier to annotate this dataset. We analyse the relationships between the temporal mental health dynamics and respective national lockdowns.

4.1 Data

We discuss the procedure for constructing the unsupervised experiment dataset, to be used for monitoring the temporal mental health dynamics.

4.1.1 Experiment Dataset

We gather tweets made public by users during the first two weeks of 2020 with a geolocation within the country of interest. We then follow the same methodology as outlined in *Section 3.1*, for the period of 1 December 2019 until 15 May 2020. The composition of these experiment datasets can be seen (*Table 9* in *Appendix A.3*) along with key dates. The key dates specified observe the official date announcements of the commencement of and of the first step towards easing of national lockdowns, rather than the first official data implementing these measures as we anticipate that the announcements would provoke users to express their opinion more than the implementation of the measures. We acknowledge caveats to the methodology with relations to the respective national lockdowns:

1. The activity-level of users whose lifestyles have been highly disrupted by the national lockdowns may be overstated during this period, due to increased leisure time.
2. The language filtering component excludes certain users of the population, such as stranded tourists/expatriates, that use a non-majority languages. Such samples may contain a bias towards a higher rate of depression.

4.2 Methodology

To monitor and analyse temporal mental health dynamics we must deploy our trained $BILSTM\text{-}SELFA$ on the respective countries' experiment datasets. Once we have the classifier's predictions, we calculate the rate of depression at any given day, R_t:

$$R_t = \frac{\sum_{i=1}^{N_t} \Phi(x_i)}{N_t} \tag{2}$$

Where Φ represents our trained classifier, x_i is the input, N_t is the total number of samples on day t. The output of the classifier, $\Phi(x_i)$ takes the form $[0, 1] \in \mathbb{N}$. R_t is a normalised continuous value between 0 and 1, interpreted as the proportion of tweets at t that classify as `Diagnosed`: 0 meaning all samples belong to `Control` and 1 meaning all samples belong to `Diagnosed`.

4.3 Results

Figures 1 and *2* (see *Appendix A.4*) display the temporal mental health dynamics for the countries under investigation. The R_t across different countries is a function of the country specific development dataset's distribution on which the classifier was trained. As such, the R_t across countries are not directly comparable but are rather analysed by the momentum of how R_t of a country changed over time and its divergence from R_t of other countries.

4.4 Discussion

Foremost, we categorically cannot, nor do we, state that the temporal mental health dynamics are *caused* by the respective national lockdowns nor other measures, taken by governments to combat the spread of the virus. In this section, we offer interpretations in line with relationships discovered.

In the U.K. rate of depression (R^{UK}), we firstly observe the sharp, unsustained, increase of over 50% on Christmas day, before decreasing back to the status quo the next day. Upon further investigation we find that this phenomenon is well-documented (Hillard & Buckman 1982) and seeing that our classifier identified this phenomenon, without explicitly being aware of its existence, is encouraging. On March 9th, Italy National Lockdown begins onward we observe a sharp, sustained increase in R^{UK} until March 23rd, U.K. National Lockdown begins, where R^{UK} somewhat plateaus. We interpret this as an increase in anxiety amongst the U.K. population as neighbouring countries take decisive measures to slow the spread of the virus. A key theme in the build up to the U.K. national lockdown implementation was the intentional delay so that to ensure maximum utility from the policy[7]. However, a report published on the 16th of March by the Imperial College COVID-19 response team[8] estimated that the the current combative approach taken up by the U.K. government would result in 250,000 deaths. The report was well-publicised by the British media and was arguably a factor in the pivot by the U.K. government.. This is somewhat supported by the change in R^{UK} during the U.K. National Lockdown where we see a sustained *decrease* over the majority of the period.

The rates of depression of France (R^{FR}), Germany (R^{DE}), Italy (R^{IT}) and Spain (R^{ES}) behave differently from R^{UK}. Firstly, R^{IT} increases sharply by over 100% in the initial days of the Italian National Lockdown. This can be interpreted as anxiety and concern as at this point Italy was regarded as the global epicentre of the pandemic. This was coupled with economic turmoil and great concern over the capacity of hospitals to handle the high requirements for intensive care units[9].

Similarly, a sharp increase in R^{FR} over the initial days of the French National Lockdown period, after-which R^{FR} rises throughout the lockdown period at a lower and inconsistent rate. A similar story could be tailored to R^{ES}. The major increase in R^{DE} occurs in the preceding month, whilst during the German National Lockdown, R^{DE} increases in the initial days, albeit at a lower rate. R^{DE} then plateaus and decreases - creating a turning point in R^{DE} during the German National Lockdown.

Furthermore, the R of respective countries following the easing of respective lockdowns can be interpreted as the countries' outlook on the easing of restrictions. The French and Spanish populations experienced a reduction in symptoms of depression, such as anxiety, as is evidenced by the clear reduction in R^{FR} and R^{ES} respectively. We therefore conclude by, tentatively, stating that the easing of restrictions were received by an improvement in the mental state of the general populations of France

[7] ITV News. 2020. Coronavirus: Boris Johnson Announces UK Government's Plan To Tackle Virus Spread, Youtube. [online] Available at: https://www.youtube.com/watch?v=2U1YoKujYeY&list=PLFXSE3NhAYiZdb2qijJ7uemIB-IAYK5-y&index=893&t=0s [Accessed 1 September 2020].

[8] Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., van Elseland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P., Walters, C., Winskill, C., Donnelly, C., Riley, Steven, R. and Ghani, A. 2020. Report 9: Impact Of Non-Pharmaceutical Interventions (Npis) To Reduce COVID-19 Mortality And Healthcare Demand. [online] Imperial.ac.uk. Available at: https://tinyurl.com/imperial-college-covid19 [Accessed 1 September 2020].

[9] CIDRAP 2020. Doctors: COVID-19 Pushing Italian ICUs Toward Collapse. [online] Available at: https://tinyurl.com/italians-covid19 [Accessed 8 August 2020].

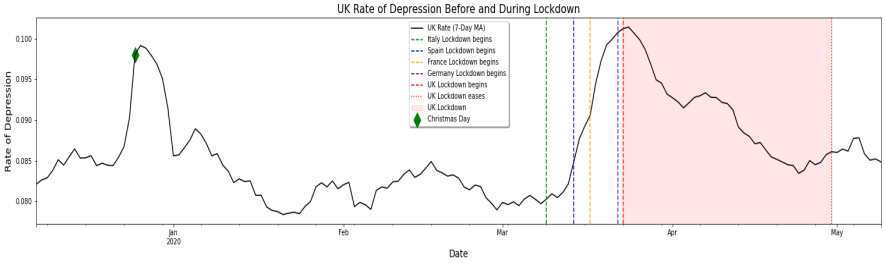


Figure 1: U.K. rate of depression before and during the National Lockdown. Noise in the rate of depression has been smoothed with a 7-day moving average.

and Spain, the mental state of the Italian and German general populations deteriorated, whilst the U.K. was agnostic to the easing of restriction.

We are hesitant to state the changes in R_t had been *caused* by the imposition/easing of national lockdowns. To make such a claim we would be required to undertake a more fine-grained causality study which is beyond the scope of this paper, however we note this for future work. We *can* however claim to have discovered clear relationships between the drastic changes in the behaviour of rates of depression during the periods of the build-up to, during and in the aftermath of national lockdowns.

4.5 Ethical Principles

As we are proposing a public data driven approach for decision-making, we offer a discussion on ethics relating to possible exploitation of the system:

One such exploitation could arise where a pharmaceutical company, focused on the anti-depressants market, utilising the methods proposed and analyse the rate of depression increasing in a particular country with no other access to an anti-depressants supplier. The company could then proceed to monopolise the market and overcharge for their products thereby constraining the individuals financially, this may in turn increase levels of stress, anxiety and depression in the country. This would create an unethical reliance on the product arising directly from the implementation of the system.

Another, yet reversed, form of exploitation could arise by taking the scenario examined in the paper, if it was publicly known that the government of a country were to utilise the methods to decide how to proceed in the easing/re- implementation of the national lockdown, it is reasonable to assume that a third-party with a vested interest in the policy setting of the government could engage in activities to manipulate the publicly available data. This could come in the form of these individuals contributing high-volume data with the sole aim to skew and corrupt the data that will be mined and used for the decision-making of the governments.

This creates a trade-off dilemma between ethical principles currently within the social media platforms user agreements stating that users have the right to know how their data is being used with the need for partial secrecy in the exact mining methodologies and their end use which lacks transparency.

A prospective equilibrium to this trade-off would be the establishment of accountable Ethics Review Boards (ERBs) at the social media network companies that will be tasked with reviewing proposed systems, judged to be too sensitive to publicly expose, developments and implementations. Furthermore, these ERBs should be audited externally periodically to ensure of their integrity. High-level details of this proposed equilibrium should be added to the social media networks' user agreements to ensure that transparency, to the extent possible, is maintained.

5 Conclusion

Our set of experiments have been conducted with the aim of providing organisations with a methodology for monitoring and analysing temporal mental health dynamics using social media data. We examine sample representations and their ability to impact classifier performance. We investigate the role of an imbalanced dataset in the classifier training regime. Our classifier achieves encouraging performance on two fronts: the ability to discriminate, with reasonable performance, between `Diagnosed` and `Control` samples and identified the Christmas Depression phenomenon. Finally, we analyse the rates of depression and their relationships with respective national lockdowns.

Ethics Board Statement

Having explicitly confirmed with the Queen Mary Ethics of Research Committee (QMERC) on a similar recent study which required Twitter data, we were advised that this type of research "does not need ethical review - being as it is analysis of data in the public domain". Additionally, we do not publicise any user IDs and/or text beyond the handful of expository examples which do not reveal any personal or identifying information. Finally, National Health Service (NHS) Research Ethics Committee (REC) review was not required for sites in England as per the online decision tool[10].

Acknowledgments

Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 program under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). 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. We express our thanks to all of our data annotators: L. Achour, L. Del Zompo, N. Fiore, M. Hechler and R. Medivil Zamudio.

References

Althoff, T., Clark, K. & Leskovec, J. (2016), 'Large-scale analysis of counselling conversations: An application of natural language processing to mental health', *TACL* .

Ayers, J., Althouse, B. & Dredze, M. (2014), 'Could behavioral medicine lead the web data revolution?', *JAMA* **311**(14), 1399–1400.

Bucci, W. & Freedman, N. (1981), 'The language of depression', *Bulletin of the Menninger Clinic* **45**(4), 334.

Chung, C. & Pennebaker, J. (2007), 'The psychological functions of function words', *Social communication* **1**, 343–359.

Coppersmith, G., Dredze, M. & Harman, C. (2014), 'Quantifying mental health signals in twitter', *In ACL Workshop on Computational Linguistics and Clinical Psychology* .

De Choudhury, M., Counts, S. & Horvitz, E. (2013*a*), 'Predicting postpartum changes in behavior and mood via social media', *In Proceedings of the SIGCHI conference on human factors in computing systems* pp. 3267–3276.

De Choudhury, M., Counts, S. & Horvitz, E. (2013*b*), 'Social media as a measurement tool of depression in populations', *In Proceedings of the Annual ACM Web Science Conference* .

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013), 'Predicting depression via social media', *In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)* .

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *In proceedings of NAACL* .

Guntuku, S., Yaden, D., Kern, M., Ungar, L. & Eichstaedt, J. (2017), 'Detecting depression and mental illness on social media: an integrative review', *Current Opinion in Behavioral Sciences* **18**, 43–49.

Hecht, H., von Zerssen, D., Krieg, C., Possl, J. & Wittchen, H. (1989), 'Anxiety and depression: comorbidity, psychopathology, and social functioning', *Comprehensive Psychiatry* **30**, 420–433.

Hillard, J. & Buckman, J. (1982), 'Christmas depression', *JAMA* **248**(23), 3175–3176.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(10), 1735–1780.

Kingma, D. & Ba, J. (2015), 'Adam: A method for stochastic optimization', *ICLR 2015* .

Li, I., Li, Y., Li, T., Alvarez-Napagao, S. & Garcia, D. (2020), 'What are we depressed about when we talk about covid19: Mental health analysis on tweets using natural language processing', *arXiv preprint arXiv:2004.10899* .

Losada, D. & Crestani, F. (2016), 'A test collection for research on depression and language use', *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016* pp. 28–39.

Oxman, T., Rosenberg, S. & Tucker, G. (1982), 'The language of paranoia', *American J. Psychiatry* **139**, 275–1282.

Paul, M. J. & Dredze, M. (2011), 'You are what you tweet: Analyzing twitter for public health', *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)* **20**, 265–272.

Peng, Z., Hu, Q. & Dang, J. (2019), 'Multi-kernel svm based depression recognition using social media data', *International Journal of Machine Learning and Cybersecurity* **10**, 43–57.

[10] NHS REC decision tool: http://www.hra-decisiontools.org.uk/ethics/index.html

Pennebaker, J., Mehl, M. & Niederhoffer, K. (2003), 'Psychological aspects of natural language use: Our words, ourselves', *Annual Review of Psychology* **54**(1), 547–577.

Rickels, K. & Schweizer, E. (1993), 'The treatment of generalized anxiety disorder in patients with depressive symptomatology', *Journal Clinical Psychiatry* **54**, 20–23.

Righetti-Veltema, M., Conne-Perréard, E., Bousquet, A. & Manzano, J. (1998), 'Risk factors and predictive signs of postpartum depression', *Journal of Affective Disorders* **49**(3), 167–180.

Rude, S., Gortner, E. & Pennebaker, J. (2004), 'Language use of depressed and depression vulnerable college students', *Cognition & Emotion* **18**(8), 1121–1133.

Shen, G., Jia, J., Nie, L., Feng, F., C. Zhang, T. H., Chua, T.-S. & Zhu, W. (2017), 'Depression detection via harvesting social media: A multimodal dictionary learning solution', *in Proceedings of IJCAI* pp. 3838–3844.

Tausczik, Y. & Pennebaker, J. (2010), 'The psychological meaning of words: Liwc and computerized text analysis methods', *Journal of Language and Social Psychology* **29**(1), 24–54.

Trotzek, M., Koitka, S. & Friedrich, C. (2018), 'Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences', *arXiv preprint arXiv:1804.07000* .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need', *In Advances in Neural Information Processing Systems* pp. 6000–6010.

Watson, W. (1768), 'An account of a series of experiments, instituted with a view of ascertaining the most successful method of inoculating the small-pox', *London: J. Nourse* .

Zhou, J., Zogan, H., S.Yang, Jameel, S., Xu, G. & Chen, F. (2020), 'Detecting community depression dynamics due to covid-19 pandemic in australia', *arXiv preprint arXiv:2007.02325* .

A Appendix

A.1 Classifier Results on All Countries' Development Datasets

Table 7: $BILSTM\text{-}SELFA$ Classifier Performance on Countries' Development Datasets

Country	Control			Diagnosed			Macro
	P	R	F1	P	R	F1	F1
France	0.96	0.94	0.95	0.29	0.37	0.33	0.64
Germany	0.96	0.96	0.96	0.3	0.32	0.31	0.63
Italy	0.97	0.96	0.97	0.38	0.4	0.39	0.68
Spain	0.97	0.97	0.97	0.35	0.38	0.36	0.67
U.K.	0.94	0.94	0.94	0.32	0.31	0.32	0.63

A.2 Significance of Results

Table 8: Significance in Predictions of $BILSTM\text{-}SELFA$ Classifier

Country	Comparison	Significance
France	Uniform	$\chi^2 = 1,311,459$ (p <0.00001)
	Weighted	$\chi^2 = 6,726$ (p <0.00001)
Germany	Uniform	$\chi^2 = 1,504,290$ (p <0.00001)
	Weighted	$\chi^2 = 5,607$ (p <0.00001)
Italy	Uniform	$\chi^2 = 1,485,204$ (p <0.00001)
	Weighted	$\chi^2 = 5,050$ (p <0.00001)
Spain	Uniform	$\chi^2 = 2,122,253$ (p <0.00001)
	Weighted	$\chi^2 = 6,324$ (p <0.00001)
U.K.	Uniform	$\chi^2 = 1,242,848$ (p <0.00001)
	Weighted	$\chi^2 = 16,591$ (p <0.00001)

A.3 Experiment Dataset Composition

Table 9: Composition of Experiment Datasets

Country	Restrictions Begin	Restrictions Eased	No. Users	No. Tweets
France	17 March 2020[1a]	11 May 2020[2a]	1,351	945,919
Germany	22 March 2020[3a]	6 May 2020[4a]	1,643	998,248
Italy	9 March 2020[5a]	27 April 2020[6a]	1,725	764,089
Spain	14 March 2020[7a]	28 April 2020[8a]	2,060	1,012,847
U.K.	23 March 2020[9a]	30 April 2020[10a]	2,883	2,050,554

[1a] The Independent. 2020. France Imposes 15-Day Lockdown And Mobilises 100,000 Police To Enforce Coronavirus Restrictions. [online] Available at: https://tinyurl.com/independent-covid19-france [Accessed 12 July 2020].

[2a] BBC News. 2020. France Eases Lockdown After Eight Weeks. [online] Available at: https://www.bbc.co.uk/news/world-europe-52615733 [Accessed 12 July 2020].

[3a] BBC News. 2020. Germany Bans Groups Of More Than Two To Curb Virus. [online] Available at: https://www.bbc.co.uk/news/world-europe-51999080 [Accessed 4 August 2020].

[4a] BBC News. 2020. Germany Says Football Can Resume And Shops Reopen. [online] Available at: https://www.bbc.co.uk/news/world-europe-52557718 [Accessed 4 August 2020].

[5a] CNN. 2020. All Of Italy Is In Lockdown As Coronavirus Cases Rise. [online] Available at: https://edition.cnn.com/2020/03/09/europe/coronavirus-italy-lockdown-intl/index.html [Accessed 12 July 2020].

[6a] BBC News. 2020. Coronavirus: Italy's PM Outlines Lockdown Easing Measures. [online] Available at: https://www.bbc.com/news/amp/world-europe-52435273 [Accessed 12 July 2020].

[7a] The Guardian. 2020. Spain Orders Nationwide Lockdown To Battle Coronavirus. [online] Available at: https://tinyurl.com/guardian-covid19-spain [Accessed 12 July 2020].

[8a] BBC News. 2020. Spain Plans Return To 'New Normal' By End Of June. [online] Available at: https://www.bbc.co.uk/news/world-europe-52459034 [Accessed 12 July 2020].

[9a] BBC News. 2020. Coronavirus Updates: 'You Must Stay At Home' UK Public Told - BBC News. [online] Available at: https://www.bbc.co.uk/news/live/world-52000039 [Accessed 12 July 2020].

[10a] BBC News. 2020. UK Past The Peak Of Coronavirus, Says PM. [online] Available at: https://www.bbc.co.uk/news/uk-52493500 [Accessed 12 July 2020].

A.4 Temporal Mental Health Dynamics Results

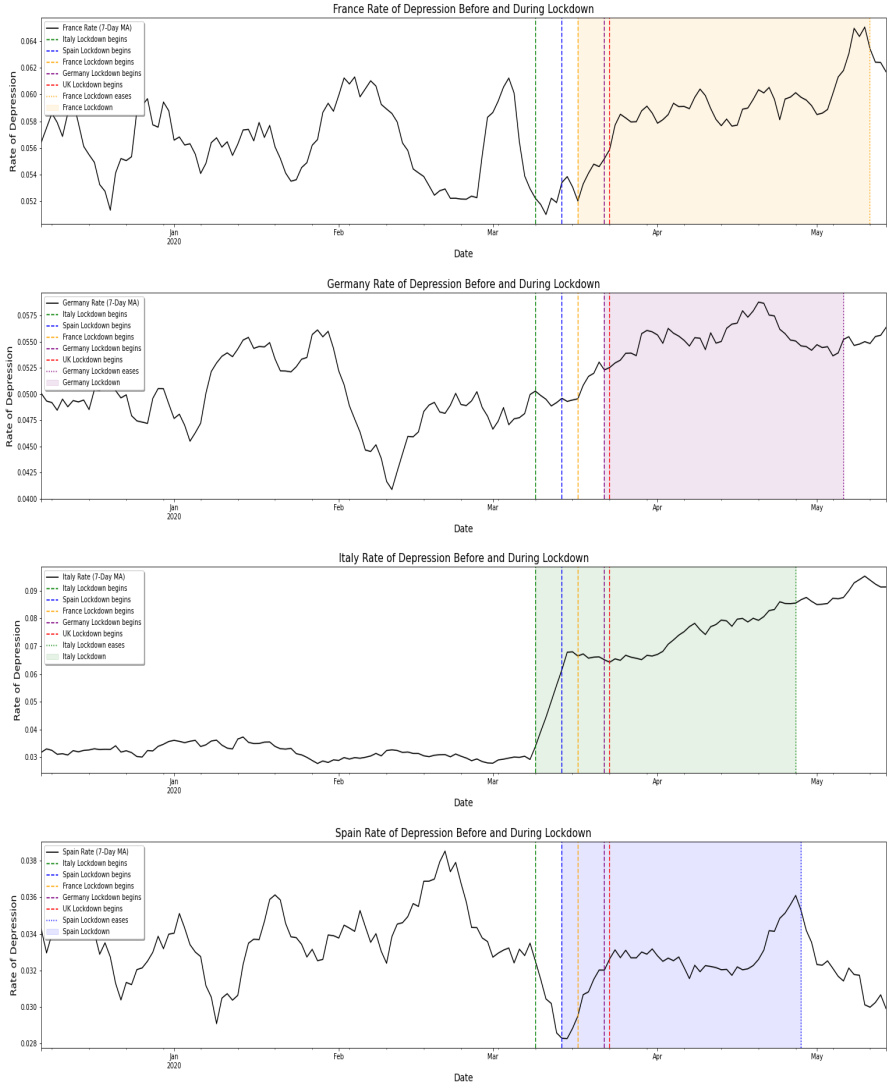


Figure 2: France, Germany, Italy and Spain rates of depression before and during respective national lockdowns. Noise in rates of depression have been smoothed with 7-day moving averages

Appendix G: Detecting and Explaining Viewpoints in Context

SemEval Task Proposal: Detecting and Explaining Viewpoints in Context

Anonymous ACL-IJCNLP submission

1 Overview

Summary: We propose a task of *detecting viewpoints in context*, and producing *explanations of system output*, for news comment data in two languages. Participants will be presented with comments posted under news articles, structured as threads linking each comment with its antecedent. In four progressively more challenging subtasks, systems must (a) detect the overall level of agreement (or disagreement) between comment and antecedent; (b) detect the issues on which the comment expresses a stance; (c) detect the polarity and strength of stance on each issue; and (d) generate a text explanation for its decisions. To avoid restricting the techniques to English, two datasets will be used, one English and one Croatian.

Motivation: Within the general field of sentiment analysis and opinion mining, most research takes the task to be one of text classification: determining the overall tone or stance of a text, with respect to some task-specific or domain-specific criteria (positive or negative opinion; author's emotional state; financial market outlook; etc.) Some tasks, however, are more focused, requiring stance towards some specific aspect or target – including recent SemEval tasks in which the stance of a text towards some given topic must be predicted (Rosenthal et al., 2017). However, although discussion between users about their stances towards given subjects is one of the primary uses of online forums and comment sections, there is little research so far that examines stances within an interactive context. Some recent examples of such research are (Zubiaga et al., 2018; Kumar and Carley, 2019) with a more detailed overview given in (Küçük and Can, 2020). On the other hand, while work in dialogue modelling often examines the interactive nature of agreement and disagreement between users, little of that examines how (dis)agreement comes together with the expression of stance. Here, we therefore propose a task and supporting dataset to encourage models which combine interaction structure with opinion mining: detecting stance towards the topics expressed in online discussion threads, together with the agreement expressed towards other commenters.

Some recent datasets and tasks do take some interactive context into consideration. The Internet Argument Corpus 2.0 (Abbott et al., 2016), for example, uses dialogue structure to annotate multiple phenomena like topic, stance, and agreement between comments. Similarly, Allaway and McKeown (2020) propose a dataset in which comments from the New York Times (NYT) 'Room for Debate' are annotated with topics and stance. However, these approaches are designed to examine argument and thus focus on explicitly *controversial* topics like birth control, where strongly polarised stance is common. However, people frequently take and express stance on *non-controversial* topics. To investigate this, we use no specific topic-based data filtering, and consider stance towards more generic topics, using a richer annotation.

To capture a wider range of phenomena, we provide richer annotations including (dis-)agreement, target, stance direction and strength, and explanation of the annotator's decision. As there are few similar datasets for non-English languages (although some examples exist, see Bošnjak and Karan, 2019; Vamvas and Sennrich, 2020), we also provide annotated data in Croatian, to encourage work for lower resourced languages or on cross-lingual approaches. We find additional motivation in recent findings on model explainability (summarized by Wiegreffe and Marasović (2021)), showing that explanations often require only shallow understanding of comments and no reasoning. This task is envisioned to require fine-grained complex analysis to generate explanations, gauging the potential

Comment	Antecedent	Text	Target	Stance	Agreement
1	N/A	I like guns and candy	[guns,candy]	[+2,+2]	N/A
2	1	I disagree! guns are bad	[guns]	[-2]	[-1]
3	1	guns are OK, candy is bad	[guns,candy]	[+2,-1]	[0]
4	3	yeah I hate it too, too sweet.	[candy]	[-2]	[+1]

Figure 1: An invented example to illustrate the data annotation schema. We omit real examples and explanations for reasons of space. Please note that, in comment 4, the target word is not mentioned.

of state-of-the-art NLP models.

Expected Impact: The rising interest in exploring explainable models makes this task very timely; conversely, the maturity of the field of sentiment and stance detection provide a context in which a challenging task, going beyond isolated text analysis to include context and explanation, is appropriate; and in which a range of models is available that are suitable for extension and application to this more challenging task. By basing our task around short texts (news comments) and simple distinctions (agreement/disagreement, positive/negative stance) we ensure that the basic parts of the task can be approached by a wide range of teams; and by providing more challenging aspects we hope to stimulate genuine progress in interactive discourse modelling and explanation generation.

2 Data and Resources

Dataset Source We use comments from two newspapers, the New York Times (NYT) in English and 24sata in Croatian; both are publicly available (see below), and we have already collated and preprocessed them as part of an ongoing project. In both cases, comments are linked to articles being commented on, and are threaded: most comments reply to a previous comment. The conversational context that emerges in threads can be key to understanding comments. We intend analysis of this context to be an important part of the task. We will randomly select 50 articles with equal distribution across a range of news categories (articles are tagged as e.g. *Politics*, *Sport*, *Finance* etc.). This ensures that our data is as diverse as possible in terms of vocabulary and topics discussed.

Annotation Annotators were recruited directly, in preference to crowdsourcing, to ensure quality in a relatively complex annotation task. We have already recruited native English and Croatian speakers in London and Zagreb, respectively; all annotators are graduate or higher-level students and paid hourly. An hour of training and 2-3 pilot annotations are used to provide feedback to ensure consistency across annotators

The annotation is divided into three phases; in the first phase, we collect the (dis)agreement between each comments and its antecedent, and a free-text explanation of this decision . In the second phase, we collect the thread's stance focus, and in the next phase, stance direction and strength towards each focus. All annotations are performed one thread at a time, with annotator first shown the whole thread to ensure the context is understood.

Resources and Availability The NYT data is collected using the NYT API[1]. For the 24sata data, we use the dataset publicly released as part of the EMBEDDIA project[2], now available on CLARIN (Pollak et al., 2021). Both datasets allow for public distribution for research and non-commercial use.

3 Proposed Tasks

Overall, the task is to detect the viewpoints expressed, characterized by their stance with regard to the target(s) of opinion (a specific topic under discussion), and by their stance towards the viewpoint(s) of the commenter being responded to. We divide this complex task into a series of subtasks.

Task A: (Dis-)Agreement Classification The simplest version of the task is to classify a given comment as agreeing or disagreeing with its antecedent comment. This can be framed as a classification task over pairs of comments: for any pair, predict the correct label from a three-way choice (agree, disagree, none/mixed). The gold-standard labels will be inferred directly from our more structured annotation (Figure 1): comments with all positive "agreement" tags are labelled *agree*; comments with all negative tags are labelled *disagree*; others are labelled *none/mixed*. Our intention with this subtask is to provide an easily-approachable version that can be suitable for a range of common classifier models.

[1]https://developer.nytimes.com/apis
[2]http://embeddia.eu/

Task B: Target Identification A more challenging task is then to predict for each comment a list of targets for which stance is being expressed (the *stance focus*, see Kiesling et al., 2018). Targets are often not mentioned in every comment, but instead must be inferred from the context: in most cases, we expect approaches that choose a key word or phrase from a comment somewhere in the thread will be able to do well, but in some cases only more adventurous approaches that generate candidate phrases or choose them from the associated news article will be able to succeed. For evaluation, we will provide all possible target lists of the dataset. Please note that a comment may have more than one target - see Figure 1.

Task C: Target-based Stance Identification Task B will then be followed by a task of identifying the stance direction and strength as regards each target, on a 5 point scale from strongly negative to strongly positive. Tasks B and C will have staggered submission deadlines in order to make evaluation and comparison of systems more direct: once entries have been closed for Task B, participants in Task C will be provided with the gold-standard lists of targets, for which stance must then be predicted.[3] Task C is therefore similar to the target-based sentiment analysis task of SemEval2017 Task 4, subtask B (Rosenthal et al., 2017) in which systems had to predict sentiment of Twitter posts towards a given topic. Here, though, there is additional information for systems to use from the thread context and author history.

Task D: Explanation Generation The final task is generative: for each classification decision in Task A, systems must produce a short text explanation of their decision. This text is expected to include the key words/phrases in the comments that make the viewpoints and (dis)agreements clear, but may rephrase or reformulate them freely. While this can be seen as a summarisation task (and we expect it to be approached by extractive and abstractive summarisation methods), it is distinct from the standard summarisation task as it must focus only on the parts of the comments which relate to the (dis)agreed-upon topics.

3.1 Pilot Tasks

We have run multiple pilots for the Agreement/disagreement with explanation tasks. The initial version of the pilot showed only the two comments in a thread and another without explanation. Also, the shuffling of comments order is tried. However, we found out that annotators preferred to read the whole thread first to get the overall context. After that, we gradually added one comment in the thread and asked the annotated agreement relationship with its antecedent. Annotators also preferred to provide an explanation of the decision on the same screen. This reduced the work needed to reread the thread. We also found out that the commentator's details also helped in the decision. Based on our pilot study of the 10 NYT threads, we found a high correlation between the agreement task and the agreement's explanation containing diverse vocabulary.

[3] Although we will also consider running Tasks B and C as a joint, open-domain target-plus-stance prediction task.

4 Evaluation

We will allow participants to submit in any tasks and in any language. For the classification tasks (Task A and C), we will use macro F1 score. As Task B is a multi-label task, we will use recall and mean average precision. For Task D, we will use both automatic metric and human evaluation: a first stage of automatic evaluation, with learned metrics like BERTScore (Zhang et al., 2019), followed by human evaluation for the top N submissions based on the automatic score.

Baseline We will provide starter code for a range of baseline models using both traditional and deep learning approaches.

5 Task Organizers

Ravi Shekhar, **Mladen Karan** and **Matthew Purver** (Queen Mary University of London) will act as main organisers, and lead dataset collection and evaluation. They have expertise in NLP for user-generated content analysis, including for stance and news comments (Shekhar et al., 2020; Bošnjak and Karan, 2019). **Andraž Pelicon**, **Senja Pollak** (Jozef Stefan Institute) and **Aleš Žagar**, **Marko Robnik-Sikonja** (University of Ljubljana) will lead the baseline creation and Croatian data verification process. All have expertise in NLP and ML and are members of the EMBEDDIA project. MP, SP and MRS were organisers of the SemEval-2020 Task 3: Graded Word Similarity in Context (Armendariz et al., 2020).

Email: {*r.shekhar, m.purver*}*@qmul.ac.uk*

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Mihaela Bošnjak and Mladen Karan. 2019. Data set for stance and sentiment analysis from user comments on Croatian news. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 50–55, Florence, Italy. Association for Computational Linguistics.

Scott F Kiesling, Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han, and Jacob Eisenstein. 2018. Interactional stancetaking in online forums. *Computational Linguistics*, 44(4):683–718.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Sumeet Kumar and Kathleen M Carley. 2019. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058.

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 Freiental, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podečan, 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*, pages 99–109.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in Croatian and Estonian. *Journal of Language Technology and Computational Linguistics*, 34:49–79. Special Issue on Offensive Language.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

Appendix H: Multi-modal Fusion with Gating using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's Dementia recognition from spontaneous speech

Morteza Rohanian[1], *Julian Hough*[1], *Matthew Purver*[1,2]

[1]Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
[2]Department of Knowledge Technologies, Jožef Stefan Institute
{m.rohanian, j.hough, m.purver}@qmul.ac.uk

Abstract

This paper is a submission to the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, which aims to develop methods that can assist in the automated prediction of severity of Alzheimer's Disease from speech data. We focus on acoustic and natural language features for cognitive impairment detection in spontaneous speech in the context of Alzheimer's Disease Diagnosis and the mini-mental state examination (MMSE) score prediction. We proposed a model that obtains unimodal decisions from different LSTMs, one for each modality of text and audio, and then combines them using a gating mechanism for the final prediction. We focused on sequential modelling of text and audio and investigated whether the disfluencies present in individuals' speech relate to the extent of their cognitive impairment. Our results show that the proposed classification and regression schemes obtain very promising results on both development and test sets. This suggests Alzheimer's Disease can be detected successfully with sequence modeling of the speech data of medical sessions.

Index Terms: Cognitive Decline Detection, Affective Computing, Computational Paralinguistics

1. Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative condition and the most common form of dementia. AD gradually affects the memory, language and cognitive skills and ultimately the ability to perform basic tasks in the everyday lives of patients. Early diagnosis of AD has become essential in disease management as it has not been possible to reverse the degenerative process, even with significant efforts focused on therapies [1].

Discrepancies in speech comprehension, speech production and memory functions are closely tied in with AD as suggested by a decrease in global vocabulary and a loss in evocative memory [2]. Patients with AD have difficulty performing tasks that leverage semantic information; they exhibit problems with verbal fluency and identification of objects [3]. The semantics and pragmatics of their language appear affected throughout the entire span of the disease more than syntax [4]. AD Patients talk more gradually with longer pauses and invest extra time seeking the right word, which contributes to disfluency of speech [3].

AD diagnosis demands the existence of cognitive dysfunction to be validated by neuropsychological assessments like the mini mental state examination (MMSE) performed in medical clinics [5]. Diagnosis is typically based on the clinical analysis of patients' history and the presence of typical neurological and neuropsychological features. It is costly and not accessible to all patients who have concerns about their memory functions.

Recent experimental research has looked at AD's automated analysis from multimodal data as alternative, less invasive tools for diagnostics. Studying behaviours of individuals could also help detect AD earlier. There has been research on building systems which use a broad range of multimodal features to identify AD severity. A meaningful association between MMSE scores and language measures such as articulation and disfluency has been found [6].

Much of the work to date has looked separately at the properties of the language of an individual: acoustic and lexical characteristics of speech, or syntax, fluency, and content of information. Usually these are studied within language tasks in specific domains or in conversational dialogue [7]. Several studies have suggested various forms of speech analysis to identify AD. Researchers found that the number of pauses, pause proportion, phonation time, phonation–to-time ratio, speach rate, articulation rate, and noise-to-harmonic ratio correlate with the severity of AD [8]. Weiner et al. [9] developed a Linear Discriminant Analysis (LDA) classifier with a set of acoustic features such as the mean of silent segments, speech and silence durations and silence to speech ratio to distinguish subjects with AD from the control group and achieved a classification accuracy of 85.7 percent. Ambrosini et al. [10] showed an accuracy of 73 percent when using selected acoustic features (pitch, voice breaks, shimmer, speech rate, syllable duration) to detect mild cognitive impairment from a spontaneous speech task.

In terms of the features which aid AD detection, lexical features from spontaneous speech are shown to be informative. Jarrold et al. [11] extracted the frequency occurrence of 14 different part of speech features and combined them with acoustic features. Abel et al. [12] modeled patient speech errors (naming and repetition disorders) to the problem of AD diagnosis.

There has also been work on modelling multimodal input for AD detection. Gosztolya et al. [13] examined the fusion of two SVM models with separate feature sets. The first model used a set of acoustic features, and the second model was developed using linguistic features extracted from manually annotated transcripts. Their work showed the complementary information that audio and lexical features may contain about a subject with AD.

Among other similar tasks, using multimodal fusion to predict a cognitive state, research has been done on integrating temporal information from two or more modalities in a recurrent approaches to classify emotions or detecting different mental states, such as depression [14]. One key challenge these mod-

http://dx.doi.org/10.21437/Interspeech.2020-2724

els have is addressing the various predictive capacity of each modality and their different levels of noise. The application of a gating mechanism in various multimodal tasks has been shown to be successful in controlling the level of contribution of each modality to the eventual prediction.

This paper addresses AD classification and MMSE score regression tasks, which are part of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [15]. In ADReSS, participants are required to assess the AD severity of different subjects, where the target severity is based on their MMSE scores.

We performed a binary classification of samples of speech into AD and non-AD classes and create regression models to predict MMSE scores. Using the ADReSS Challenge data which consists of speech recordings and transcripts of spoken picture descriptions, we explored various features as diagnostically relevant tools. We focused in particular on sequential modelling of sessions and whether the disfluencies and self-repairs present in individuals' speech can help predict the level of cognitive impairment.

Our approach is motivated by [14] that developed the ability to learn difficult decision boundaries which other models with different methods of fusion have trouble managing, and maximise the use and combination of each modality. We employed data of individuals under controlled conditions, and modeled the sessions with audio and text features in a Long-Short Term Memory (LSTM) neural network to detect AD. Our findings indicate that AD can be detected with minimal information available on the structure of the description tasks by pure sequential modelling of a session. We also found that disfluency markers have predictive power for AD recognition.

2. Proposed Approach

Our approach is to model the speech of individuals giving picture descriptions as a sequence to predict whether they have AD or not, and if so, to what degree. To predict AD, we performed three sets of experiments using features from the audio and text data:

1 LSTM models utilising unimodal audio and text features.

2 LSTM model with gating to test the effect of using multimodality.

3 A multimodal LSTM model using acoustic and lexical information, including disfluency tagging.

The details of the three experiments are outlined below in the following sub-sections. In line with the standard assumption in deep learning, we take the approach that for a model to be genuinely data-driven, minimal feature engineering is required. The model's power is in its capacity to represent information through non-linear transforms, at varying spatial and temporal units, and from different modalities. Since we were interested in modelling temporal session changes, we used a bi-directional Long Short-Term Memory (LSTM) neural network as it has the added benefit of sequential data modelling. For each of the audio and text modalities we trained an LSTM model separately, using the audio and text features.

2.1. Multimodal Features

Lexical Features from Text A pre-trained GloVe model [16] was used to extract the lexical feature representations from the picture description transcript and convert the utterance sequences into word vectors. We selected the hyperparameter values, which optimised the output of the model on the training set. The optimal dimension of the embedding was found to be 100.

Audio Features A set of 79 audio features were extracted using the COVAREP acoustic analysis framework software, a package used for automatic extraction of features from speech [17]. We sampled the audio features at 100Hz and used the higher-order statistics (mean, maximum, minimum, median, standard deviation, skew, and kurtosis) of COVAREP features. The features include prosodic features (fundamental frequency and voicing), voice quality features (normalized amplitude quotient, quasi open quotient, the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum, maxima dispersion quotient, parabolic spectral parameter, spectral tilt/slope of wavelet responses, and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics) and spectral features (Mel cepstral coefficients 0-24, Harmonic Model and Phase Distortion mean 0-24 and deviations 0-12). Segments without audio data were set to zero. A standard zero-mean and variance normalization was applied to features. We omitted all features with no statistically significant univariate correlation with the results of training set.

2.2. Sequence Modeling

The potential of neural networks lies in the power to derive representations of features by non-linear input data transformations, providing greater power than traditional models. As we were interested in modelling temporal nature of speech recordings and transcripts, we used a bi-directional LSTM. For each of the audio and text modalities we trained a separate unimodal LSTM model, using different sets of features. For the input data we explored different timesteps and strides. After exploring different hyper-parameters, the model using audio data has a timestep of 20 and stride of 1 with 4 bi-directional LSTM layers with 256 hidden nodes. The model using text input has an input with a timestep of 10 and stride of 2 and has 2 LSTM layers with 16 hidden nodes. The code used in the experiments are publicly available in an online repository.[1]

2.3. Multimodal Fusion with Gating

Audio and text features can include not only discriminative and temporarily changing information about the current state of a subject, but supporting information as well.

The model consists of two branches of the LSTM, one for each of the modalities, with their outputs combined into final feed-forward highway layers. The branches are made up of different hyperparameters and configured with respect to each modality's properties. Their outputs are concatenated and passed through N highway layers (where the best value N was determined from optimizing on heldout data). We pad the size of the training examples in the text set (which was the smaller set) to meet the audio set by mapping together instances that occurred in the same session, as the audio and text inputs for each branch of the LSTM had different timesteps and strides.

Gating Mechanism Data from two modalities affect the final output differently, and it is important to consider the amount of noise when aggregating them into a single representation. Since learned representation for the text can be undermined by corresponding audio representation, during multimodal fusion we need to minimise the effects of noise and overlaps. We use feed-forward highway layers [18], with gating units that learn by weighing text and audio inputs at each time step to regulate

[1] https://github.com/mortezaro/ad-recognition-from-speech

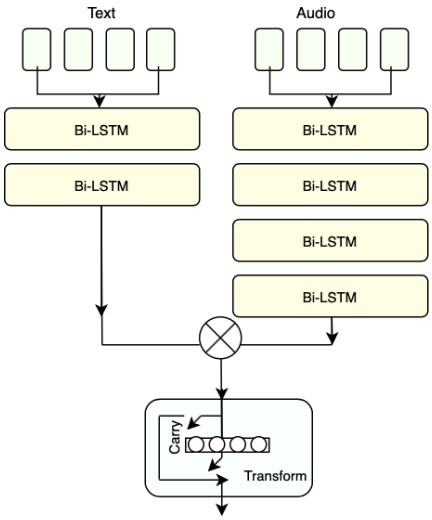


Figure 1: *Multimodal fusion with gating.*

information flow through network work.

Each highway layer consists of two non-linear transformations: a Carry (Cr) and a Transform (Tr) gate which determine the degree to which the output is generated by transforming and carrying the input. Each layer uses the gates and feed-forward layer H to regulate its input vector at timestep t, D_t, to generate output y:

$$y = Tr \cdot H + Cr \cdot D_t \quad (1)$$

where Cr is simply defined as $1 - Tr$, giving:

$$y = Tr \cdot H + (1 - Tr) \cdot D_t \quad (2)$$

The transform gate Tr is defined as $\sigma(W_{Tr}D_t + b_{Tr})$, where W_{Tr} is the weight matrix and b_{Tr} the bias vector for the gates. Based on the transform gates outputs, highway layers adjusts their performance from multiple-unit layers to layers that only pass through their inputs. As inspired by [18] and to help resolve long-term learning dependencies faster we initialise b_{Tr} with a negative value (biased towards the Carry gate). We use a block of 3 stacked highway layers. The overall architecture of the LSTM with Gating model is shown in Figure 1.

2.4. Multi-modal Model with Disfluency Markers

Disfluencies like self-repairs, pauses and fillers are widespread in everyday speech [19]. Disfluencies are usually seen as indicative of communication problems, caused by production or self-monitoring issues [20]. Individuals with AD are likely to deal with troubles in language and cognitive skills. Patients with AD speak more slowly and with longer breaks, and invest extra time seeking the right word, which in effect contributes to disfluency [3]. The present research explores the disfluencies present in the speech of AD patients as they contribute to severity of symptoms.

Self-repair disfluencies are typically assumed to have a reparandum-interregnum-repair structure, in their fullest form as speech repairs [21]. A reparandum is a speech error subsequently fixed by the speaker; the corrected expression is a repair. An interregnum word is a filler or a reference expression between the words of repair and reparandum, often a halting step as the speaker produces the repair, giving the structure as in (3)

$$\texttt{John} \ \underbrace{\texttt{[likes}}_{\texttt{reparandum}} + \underbrace{\texttt{\{ uh \}}}_{\texttt{interregnum}} \ \underbrace{\texttt{loves]}}_{\texttt{repair}} \ \texttt{Mary} \quad (3)$$

In the absence of reparandum and repair, the disfluency reduces to an isolated *edit term*. A marked, lexicalized edit term such as a filled pause ("uh" or "um") or more phrasal terms like "I mean" and "you know" can occur. Recognizing these elements and their structure is then the task of disfluency detection.

We automatically annotated self-repairs using a deep-learning-driven model of incremental detection of disfluency developed by Hough and Schlangen [22, 23]. It consists of deep learning sequence models that use word embeddings of incoming words, part-of-speech annotations, and other features in a left-to-right, word-by-word manner to predict disfluency tags. Here each word is either tagged as a repair onset tag (marking first word of the repair phase) edit term, or fluent word by the disfluency detector- we concatenate the disfluency tags with the word vectors to create the input for text-based LSTM.

3. Experiments

3.1. Data

The ADReSS challenge's data consists of speech recordings and transcripts of spoken picture descriptions gathered from participants via the Boston Diagnostic Aphasia Exam's Cookie Theft picture [15]. The training set includes 108 subjects, and the state of the subjects is assessed on the basis of the MMSE score. MMSE is a commonly used cognitive function test for older people. It involves orientation, memory, language, and visual-spatial skills tests. Scores of 25-30 out of 30 are considered as normal, 21-24 as mild, 10-20 as moderate and $<$10 as severe impairment.

The total number of speech segments each participant had generated was 24.86 on average. The annotations for the test set were not included in the public release of the ADReSS Challenge, so all models were tested on both the development and test set. The data is pre-processed acoustically and is balanced in terms of age and gender.

3.2. Implementation and Metrics

We set up our model to learn the most useful information from modalities for predicting AD. All experiments are carried out without being conditioned on the identity of the speaker. The sizes of layers and the learning rates are calculated by grid search on validation test. The LSTM models were trained using ADAM [24] with a learning rate of 0.0001. For the loss function we used Binary Cross-Entropy to model binary outcomes, and Mean Square Error (MSE) to model regression outcomes. For binary classification of AD and non-AD, we report accuracy, precision, recall, and F1 scores and for the MMSE prediction task we report the Root Mean Square Error (RMSE).

3.3. Baseline Models

We compare the performance of our models to the ADReSS Challenge baseline [15] with an ensemble of audio features which was provided with the dataset. The baseline classification experiments were different methods of linear discriminant

analysis (LDA), decision trees (DT), and support vector machines (SVM). The baseline regression experiments were different methods of DT, gaussian process regression (GPR), and SVM.

Table 1: *Result of the AD classification and regression experiments with our models in cross validation*

Models	Features	Accuracy	RMSE
LSTM	Acoustic	0.64	6.01
LSTM	Lexical	0.69	5.42
LSTM	Lexical+ Dis	0.73	5.08
LSTM with Gating	Acoustic + Lexical	0.76	5.01
LSTM with Gating	Acoustic + Lexical + Dis	**0.77**	**4.98**

Table 2: *Result of the AD classification and regression experiments with our models against baseline models on test set*

Models	Features	Accuracy	RMSE
Baseline ([15])			
LDA	Acoustic	0.625	-
DT	Acoustic	0.625	6.14
SVM	Acoustic	0.563	6.12
GPR	Acoustic	-	6.33
Our Models			
LSTM	Acoustic	0.666	5.93
LSTM	Lexical	0.708	5.45
LSTM	Lexical + Dis	0.729	4.88
LSTM with Gating	Acoustic + Lexical	0.771	4.57
LSTM with Gating	Acoustic + Lexical + Dis	**0.792**	**4.54**

4. Results

In Table 1, we present our proposed model's performance in a cross-validation setting and in Table 2 against that of baselines models on AD detection on the provided test set. For AD detection, our proposed LSTM model with gating and disfluency features achieves an accuracy of **0.792** and RMSE of **4.54**, outperforming all the baselines. The overall findings confirm our assumption that a model with a gating structure can more efficiently minimise the errors and noise of the individual modalities.

Effect of disfluency features We found that disfluency tags help as features in AD detection. Adding disfluency features to the lexical features lead to improvement in both unimodal (ACC 0.70 vs. 0.72; RMSE 5.45 vs. 4.88) and multimodal models (ACC 0.77 vs. 0.79; RMSE 4.57 vs. 4.54).

Effect of multimodality The multimodal LSTM with gating model outperforms the single modality AD detection models in both the classification and regression tasks. A performance increase is obtained by combining textual and audio modalities with gating over single modality models (ACC 0.72 vs. 0.79; RMSE 4.88 vs. 4.54). Adding audio features improves performance despite having different steps and timesteps inputs for each LSTM branch. In terms of our competitor baselines (without the information from the manual transcripts), multimodal classifiers performed better than all the baseline models, indicating the potential benefits of multimodal fusion in AD detection. We found that while the baseline audio-based models have some discriminative capacity, sequence modelling is more accurate (ACC scores 0.67 vs. 0.63) and has lower (better) RMSE (5.93 vs. 6.12) for predicting AD.

For AD classification, the text features alone are more informative than the audio features, as using only the text modality gives a better AD prediction than utilizing unimodal audio modality sequentially (Acc scores 0.73 vs. 0.67; RMSE 4.88 vs. 5.93).

We can see that all models provide more accurate results on the test set than in cross validation. LSTM with gating models accuracy improved more than other models on the test set (RMSE 4.54 and 4.57 vs. 4.98 and 5.01).

Error analysis The results in Table 3 show that the LSTM model with gating and disfluency features obtains the highest precision and recall for both AD and non-AD classes. The model achieves F1 scores of 0.7826 for AD and 0.8000 for non-AD. The addition of gating particularly improves the recall of AD class: the LSTM model with lexical and disfluency features without gating has a recall 0.6667 for the AD class compared to the 0.7500 achieved with gating, while its 0.7910 recall for the non-AD class is not as far beneath the 0.8333 achieved by the full gating model. Depending on the application the model is used for, false negatives or false positives for AD detection will be more or less desirable, but as it stands our full gating model considerably reduces the false negatives of diagnosis whilst still marginally reducing the false positives.

Table 3: *Results of AD classification task on test set*

Models	Class	Precision	Recall	F1 Score	Accuracy
LSTM	AD	0.7619	0.6667	0.7111	0.7292
(Lexical+ Dis)	non-AD	0.7037	0.7910	0.7451	
LSTM with Gating	AD	0.7826	0.7500	0.7660	0.7708
(Acoustic + Lexical)	non-AD	0.7600	0.7917	0.7755	
LSTM with Gating	AD	0.8182	0.7500	0.7826	0.7917
(Acoustic + Lexical+ Dis)	non-AD	0.7692	0.8333	0.8000	

5. Conclusions

We have presented a deep multi-modal fusion model that learns the AD indicators from audio and text modalities as well as disfluency features. We trained and tested the model on audio and transcript data from individuals doing a description task under controlled conditions, and modeled the sessions with an LSTM and feed-forward highway layers as gating mechanism for AD detection. Our findings indicate that AD can be identified by pure sequential modelling of a session, with limited information available on the structure of the description tasks. We also found that markers of disfluency hold predictive power for identification of AD.

In future work we intend to study a series of language markers associated with AD severity, as well as interactions between them. In particular, we want to undertake a more principled approach to lexical markers, disfluency markers in terms of a study of self-repair and structural markers with a look at grammatical fluency. Furthermore, we want to find acoustic features that contribute more to the prediction of AD and have higher correlation with linguistic information.

6. Acknowledgments

Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). 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.

7. References

[1] A. Burns and S. Iliffe, "Alzheimer's disease," *B M J*, vol. 338, no. 7692, pp. 467–471, 2 2009.

[2] D. Kempler, *Neurocognitive disorders in aging*. Sage, 2005.

[3] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[4] K. A. Bayles and D. R. Boone, "The potential of language tasks for identifying senile dementia," *Journal of Speech and Hearing Disorders*, vol. 47, no. 2, pp. 210–217, 1982.

[5] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[6] M. F. Weiner, K. E. Neubecker, M. E. Bret, and L. S. Hynan, "Language in alzheimer's disease," *The Journal of clinical psychiatry*, vol. 69, no. 8, p. 1223, 2008.

[7] S. Nasreen, M. Purver, and J. Hough, "Interaction patterns in conversations with alzheimer's patients."

[8] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[9] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[10] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid *et al.*, "Automatic speech analysis to early detect functional cognitive decline in elderly population," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 212–216.

[11] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[12] S. Abel, W. Huber, and G. S. Dell, "Connectionist diagnosis of lexical disorders in aphasia," *Aphasiology*, vol. 23, no. 11, pp. 1353–1378, 2009.

[13] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[14] M. Rohanian, J. Hough, M. Purver *et al.*, "Detecting depression with word-level multimodal fusion," *Proc. Interspeech 2019*, pp. 1443–1447, 2019.

[15] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

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

[17] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[19] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language*, vol. 53, no. 2, pp. 361–382, 1977.

[20] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[21] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, Citeseer, 1994.

[22] J. Hough and D. Schlangen, "Recurrent neural networks for incremental disfluency detection," ser. Interspeech 2015, 2015, pp. 849–853.

[23] ——, "Joint, incremental disfluency detection and utterance segmentation from speech," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 326–336.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.