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

D3.2: Initial cross-lingual context and opinion analysis technology (T3.1)

Executive summary
This deliverable summarises progress to date (M18) on Task T3.1 of the EMBEDDIA project, which 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 and T3.3. We review existing work in user-generated content analysis, mostly in the domain of social media, as few directly relevant resources exist in the news comment domain. We summarise our progress in developing new classifiers for the detection of a range of relevant aspects, including author type, non-human authors, opinion, and sentiment; and in developing new models to incorporate context into the analysis. We show that our methods are suitable for cross-lingual transfer, applying the transfer methods investigated in WP1 to apply to less-resourced EMBEDDIA languages with little performance drop.

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:	**D3.2**
Deliverable full title:	**Initial cross-lingual context and opinion analysis technology**
Deliverable short title:	**Initial cross-lingual comment analysis**
Document identifier:	**EMBEDDIA-D32-InitialCrosslingualCommentAnalysis-T31-submitted**
Lead partner short name:	**QMUL**
Report version:	**submitted**
Report submission date:	**30/06/2020**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Matthew Purver (QMUL), Ravi Shekhar (QMUL)**
Co-author(s):	
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
14/05/2020	v0.1	Matthew Purver (QMUL)	Initial draft.
25/05/2020	v1.0	Matthew Purver (QMUL)	First complete draft.
12/06/2020	v1.1	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	Submitted for internal review.
14/06/2020	v1.2	Marko Robnik-Šikonja (UL)	Internal review.
24/06/2020	v1.3	Senja Pollak (JSI)	Internal review.
25/06/2020	v1.4	Matthew Purver (QMUL), Ravi Shekhar (QMUL)	Revision based on internal reviews.
27/06/2020	v1.5	Nada Lavrač (JSI)	Report quality checked and finalised.
30/06/2020	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ELMo	Embeddings from Language Models
IRC	Internet Relay Chat
LASER	Language Agnostic SEntence Representations
LSTM	Long Short-term Memory
NER	Named Entity Recognition
NLP	Natural Language Processing
RNN	Recurrent Neural Network
UGC	User-Generated Content
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 the content 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 architecture is general enough to be applicable to a range of specific tasks and re-trained in Tasks T3.2 and T3.3. 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 LSTM-RNNs and hybrid BERT-based DNNs, which are suitable for integration with WP1 for cross-lingual transfer. We applied these models to a range of classification tasks relevant to user-generated content, including author profiling (Martinc et al., 2019b; Martinc & Pollak, 2019), bot detection (Martinc et al., 2019a), sentiment detection (Robnik-Šikonja et al., 2020), and opinion detection (Concannon & Purver, in preparation).

Work in the first phase concentrated on monolingual approaches, training on standard datasets mostly from social media. However, the use of multi-lingual encoders such as multilingual BERT and LASER embeddings means that they are suitable for the cross-lingual transfer with methods developed in WP1. Our second phase of work investigated cross-lingual transfer, training on one language and testing on another, for a number of languages including project languages Croatian, Slovene, Swedish and Russian (Robnik-Šikonja et al., 2020). This approach has now been combined with new models from WP1 and extended to comment filtering in Task T3.2 (see Deliverable D3.3).

Work in the first phase treated texts as independent; this is not the case, of course, as comments occur in the context of a news article and of the developing feed of other comments. Our ongoing work is now developing models of context that can help improve the analysis accordingly, by enabling fusion of information in different modalities, and analysing relations between texts (e.g., the agreement and disagreement relations between news comments, important for understanding and summarising opinions). Building on empirical work in question-answer relations (Nasreen et al., 2019a)(Nasreen et al., 2019b) and context fusion (Rohanian et al., 2019) we are developing classifiers for improving profiling accuracy, and detecting thread relations and agreement.

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

- Classifiers for author profiling (including gender detection) with state-of-the-art accuracy, ranked 2nd in the CLIN cross-genre author profiling shared task (Martinc & Pollak, 2019) and 3rd in the PAN author profiling shared task (Martinc et al., 2019b).

- Classifiers for bot and gender detection with good accuracy, ranked 16th in the PAN 2019 bot and gender profiling shared task (Martinc et al., 2019a).
- Classifiers for opinion detection, based on BERT models suitable for cross-lingual transfer (Concannon & Purver, in preparation).
- Classifiers for sentiment detection using LASER and BERT, and showing that these models can transfer cross-lingually (Robnik-Šikonja et al., 2020).
- Methods for incorporating contextual knowledge into DNN classifiers (Rohanian et al., 2019), for application to modelling relations of comments to articles.
- Methods for using question-answer relations in user profiling (Nasreen et al., 2019a), (Nasreen et al., 2019b), now being applied to modelling news comment relations.

This report is split into 6 further sections. Section 2 summarises related work in analysing news comments and other UGC. In Section 3, we describe our work on characterising text by aspects of its author, including detection of UGC produced not by humans but by bots. Section 4 describes our classifiers for detection of sentiment and opinion, including experiments to show cross-lingual transfer potential (training on datasets in well-resourced languages and testing on others). Section 5 summarises our ongoing work in relating meaning to context: enabling classifiers to use context, including thread structure, to improve analysis. Section 6 summarises the main concrete outputs of this work, and Section 7 summarises our conclusions and main findings, and outlines the plans for further work. 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 to be 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 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.

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 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.

2.3 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 will also be a key component: accuracy in tasks such as sentiment and opinion detection in news comments, and the comment filtering in Task T3.2, will 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 be 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.

3 Author analysis

Our first steps in this task were to develop classifiers using general architectures which could be applied to UGC analysis tasks, with the first tests being on categorising different characteristics of the author. We applied this to detection of author gender (Section 3.1) of high-profile authors (Section 3.2), and to distinguishing of automated bots from human authors (Section 3.3); these tasks provide potentially useful information for filtering (Task T3.2) and summarisation (Task T3.3).

3.1 Gender analysis

The CLIN 2019 shared task in gender classification provides data in the Dutch language in three genres: Twitter, YouTube comments, and news articles. The task requires systems to predict author gender, a task that produces information likely to be useful in T3.3 for, e.g., summarising and comparing opinions and reactions by gender. The task requires prediction both *in-genre* (training and testing on data from each genre separately) and *cross-genre* (for each genre, training only on the data from the other genres). Our entry used a deep neural network (DNN) architecture (see Figure 1) in which the input word and part-of-speech sequences are each encoded via embeddings and fed to a bidirectional LSTM layer, which is followed by average and max pooling layers. The outputs are concatenated and passed through a fully connected layer. This achieved 2nd place in the shared task for cross-genre ranking, with a best cross-genre accuracy of 55.2% when trained on Twitter data and YouTube comments, and tested on the balanced news corpus. The best in-genre accuracy of 61.33% was achieved on YouTube comments, and the system ranked 6th overall in the in-genre ranking; see Table 1 for results.

Table 1: Classification accuracy achieved on CLIN 2019 gender prediction task. Genres provided were Twitter, YouTube comments and news articles; in-genre results are for training and testing on the same genre, while cross-genre results test on one genre after training on the others two (Martinc & Pollak, 2019).

	Twitter	YouTube	News	Average
Validation set in-genre	0.6245	0.6270	**0.6477**	0.6331
Validation set cross-genre	0.5473	**0.5580**	0.5573	0.5542
Official test set in-genre	0.6099	**0.6133**	0.5990	0.6074
Official test set cross-genre	0.5427	0.5507	**0.5520**	0.5485

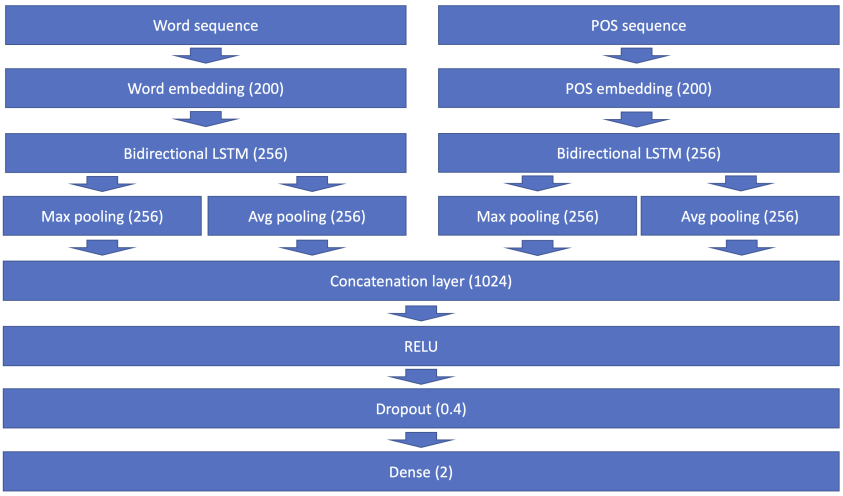

Figure 1: Network structure for gender prediction classifier (Martinc & Pollak, 2019)

The system was used in a monolingual setting here, but the approach is well suited to cross-lingual transfer, either initialising embeddings from a cross-lingually aligned static model, or replacing the standard LSTMs with a cross-lingual ELMo model (Peters et al., 2018) as developed in WP1. The success in the cross-genre task suggests the approach is sufficiently robust for this.

This work is described in full in (Martinc & Pollak, 2019), attached here as Appendix A.

3.2 User type analysis

The PAN@CLEF 2019 shared task in Celebrity Profiling provided another chance to develop and test on an author classification task. In this case, the datasets were taken from social media in English, and the task was to classify according to 4 different dimensions. Each dimension was a multi-class problem: *fame* could take the values *superstar*, *star*, *rising*; gender was defined as being one of *male*, *female*, *non-binary*; *occupation* could take 8 different values and *birthyear* 70 different values. The classes were very unevenly balanced; for example, nearly 75% of cases had the classification *fame=star* with under 5% *fame=rising*; while nearly 40% of cases had *occupation=sports* with barely 0.1% *occupation=religious*.

We tested a variety of approaches, including the use of state-of-the-art pretrained language model BERT (Devlin et al., 2019) (as used in much of the cross-lingual work in WP1), but significantly better performance on the validation set was obtained by simpler models, therefore our final submission used word and character features within a logistic regression model (see Table 2). Such a model would be suitable for a cross-lingual setting when combined with cross-lingual dictionaries or aligned static embedding spaces. The performance on the test sets was noticeably lower than on the validation set, but this was a difficult task, and our system came 3rd in the final ranking.

Table 2: Classification results achieved on CLEF 2019 Celebrity Profiling task (F1 scores for specific categories and cRank for All) (Martinc et al., 2019b).

	Fame	Gender	Occupation	Birthyear	All
Validation dataset	0.7837	0.9017	0.7578	0.0649	0.2092
Test dataset 1	0.517	0.580	0.449	0.361	0.462
Test dataset 2	0.507	0.594	0.486	0.347	0.465

This work is described in full in (Martinc et al., 2019b), attached here as Appendix B.

3.3 Bot detection

The PAN@CLEF 2019 shared task in Bots and Gender Profiling provides a testbed with a classification problem directly relevant to WP3's objectives of comment *filtering* (T3.2) as well as summarisation (T3.3) - in the filtering task, the identification of automated bots is likely to be crucial. The datasets were again taken from UGC on social media (Twitter), this time in English and Spanish, and the task was to classify in a two-step process: first detecting tweets authored by bots rather than humans, and then detecting the gender of the human authors.

For this task, we used the approach from Section 3.2 above: a linear classifier (Logistic Regression) with a range of word and character n-gram features, supplemented with a simple type-to-token-ratio feature (this proved particularly useful for bot detection). The bot detection subtask proved easier than gender detection (see Table 3). The accuracies achieved were high, with the bot detection subtask achieving 89% classification accuracy on English data, 87% on Spanish, and the system ranked 16th overall (on bot and gender prediction). It is notable that the winning entry (Pizarro, 2019), achieving over 93% accuracy for the bot detection subtask in both languages, also used a linear model (a linear kernel SVM) with similar word and character n-gram features, rather than a more complex NN model.

Table 3: Classification accuracies achieved on CLEF 2019 Bots and Gender Profiling task (Martinc et al., 2019a).

	Validation set		Test set	
	Bot	Gender	Bot	Gender
English	0.9016	0.7952	0.8939	0.7989
Spanish	0.8804	0.6696	0.8744	0.7572

The work in this section is multilingual and shows that the approaches are general enough to succeed across languages, although evaluation of cross-lingual transfer is left to later tasks; see Deliverable D3.3 for our work in cross-lingual UGC filtering, and Section 4.2 for cross-lingual sentiment analysis.

This work is described in full in (Martinc et al., 2019a), attached here as Appendix C.

4 Sentiment and opinion analysis

The second thread of UGC content analysis work has focused on developing methods for 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). Again, we begin with monolingual work, although using methods compatible with our cross-lingual transfer work in WP1 (see Section 4.1). We then show that the methods studied in WP1 can be incorporated and test our sentiment analysis approach in an explicitly cross-lingual setting (Section 4.2).

4.1 Opinion, topic and stance detection

One set of experiments focused on the detection of UGC texts that contain opinions, together with classification of the sentiment of those opinions (i.e. stance), and their categorisation by topic to allow clustering according to the issue on which an opinion is being expressed. To create a dataset, we streamed a large dataset of Twitter messages in English around a particular topic (in this case, the UK health service; the data collection originally took place as part of another project and was made available to EMBEDDIA) by following a manually defined set of search terms. We then trained a supervised classifier to distinguish opinion-containing texts from others, applied a sentiment classifier to distinguish positive from negative opinions, and used an unsupervised topic model to discover the issues on which opinions were expressed.

The dataset used was large and noisy: the applied collection method means that many off-topic tweets are collected, and that the data is highly unbalanced (only c.1-2% of data contained opinions on qual-

ity of healthcare). To derive a suitably labelled set, we used a version of *active learning*: an iterative process with several stages of classifier retraining and prediction, interleaved with manual correction of predicted-positive examples. Given a suitable training set, we compared a linear-kernel SVM with word n-gram features, and a fine-tuned BERT model. BERT outperformed the SVM, giving over 70% macro-averaged F_1 score, with 48.5% F_1 on the task of detecting the (minority) opinion-containing class (vs. 37.7% F_1 with SVM).

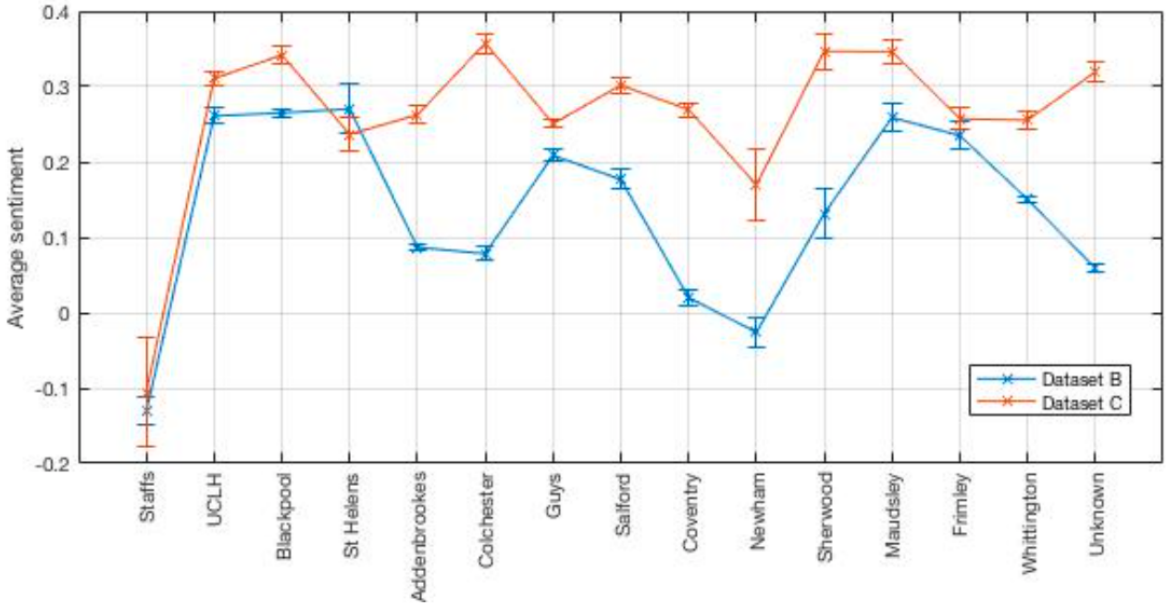

Figure 2: Variation of opinion sentiment by target (specific hospital), and over time (Dataset B collected 06/09/2015–22/02/2016, Dataset C 23/02/2016–29/09/2016) (Concannon & Purver, in preparation).

Sentiment classification was performed using a linear-kernel SVM trained on a separate general English Twitter dataset, using distant supervision to label the positive sentiment based on the presence of emoticons and hashtags (Purver & Battersby, 2012). Accuracy was good, with 78.2% F_1 score on negative opinions and 76.5% on positive opinions (outperforming a dictionary-based baseline and a DNN approach trained on newspaper text). The resulting outputs allow analysis of how the sentiment of opinions varies over time and by target (see Figure 2). Results also showed that more negative opinions were discovered using our BERT-based opinion classifier than when using hashtag and account name-based methods, as used in related work (Greaves et al., 2014; Hawkins et al., 2016).

1	london, whilst, building, disregard, contempt, utter, food
2	years, life, night, free, saving, ago, cancer
3	care, staff, family, nursing, service, incredible, brilliant
4	bad, paramedics, treatment, part, poor, busy, worse
5	a&e, waiting, failed, victoria, feel, money, times
6	staff, excellent, amazing, shows, ceo, whilst, building
7	care, great, staff, patient, amazing, hope, area

Table 4: Example topics derived using LDA: first 7 topics with top 7 keywords for each (Concannon & Purver, in preparation).

Finally, topic was categorised first by deriving a thematic list by manual inspection, and then using LDA (Blei et al., 2003) to infer topics in an unsupervised manner. The manual analysis showed that the opinions detected covered a range of important topics: as well as the major category of patient care, opinions also related to themes such as quality of hospital environment, transport access and waiting times. Sentiment analysis showed that opinions on care quality were the least negative, with those on

environment, transport and waiting times the most negative. Comparison of the automatically-inferred topic word lists showed that several matched manually identified themes (patient care, staff quality, waiting times), although many were harder to interpret and tended to mix sentiment indicators with topical content; see Table 4 for examples.

This work therefore provides several of the main building blocks required for summarisation in Task T3.3, but is so far monolingual, tested only on English. Nevertheless, Using active learning, a pre-trained sentiment classifier based on distant supervision, and an unsupervised topic model, it is well suited to transfer to other domains and languages with little or no annotated data.

This work is described in full in (Concannon & Purver, in preparation); in order to maintain anonymity for review, this is not attached here but can be made available on request and will be included as part of Deliverable D3.4.

4.2 Sentiment detection and cross-lingual transfer

In the subtask of sentiment analysis, we examined the use of multi-lingual pre-trained models and transfer learning to allow cross-lingual transfer. In this work, we again used UGC text from Twitter data; here we consider the positive/neutral/negative sentiment split as a single three-way classification task (rather than the two-stage opinion/neutral and positive/negative pipeline used in Section 4.1 above). We used a dataset with over 1.6 million manually annotated tweets in 15 languages (Mozetič et al., 2016), experimenting with 13 of those languages which showed best inter-annotator agreement.

Our classification method was based on the use of multi-lingual sentence encoders, which project texts into sentence embeddings in a shared multi-lingual space. We compared the use of LASER (Artetxe & Schwenk, 2019), passing the outputs of the LASER LSTM models through a densely-connected NN layer, and BERT (Devlin et al., 2019), fine-tuning the last layer of its Transformer stack.

Table 5: Classifier performance as macro-averaged F_1 score and classification accuracy (CA) for the three-way sentiment task with cross-lingual transfer using LASER: training on *source* language data and testing on *target* language data. The drop in performance compared to the ideal monolingual case is less with transfer between languages in the same family (Table (a)) than different families (Table (b)) (Robnik-Šikonja et al., 2020).

(a)

		Transfer		Both target	
Source	Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
German	English	0.55	0.59	0.62	0.65
Polish	Russian	0.64	0.59	0.70	0.70
Polish	Slovak	0.63	0.59	0.72	0.72
German	Swedish	0.58	0.57	0.67	0.65
German Swedish	English	0.58	0.60	0.62	0.65
Slovene Serbian	Russian	0.53	0.55	0.70	0.70
Slovene Serbian	Slovak	0.59	0.52	0.72	0.72
Average performance gap		0.09	0.11		

(b)

		Transfer		Both target	
Source	Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
Russian	English	0.52	0.56	0.62	0.65
English	Russian	0.57	0.58	0.70	0.70
English	Slovak	0.46	0.44	0.72	0.72
Polish, Slovene	English	0.58	0.57	0.62	0.65
German, Swedish	Russian	0.61	0.61	0.70	0.70
English, German	Slovak	0.50	0.47	0.72	0.72
Average performance gap		0.14	0.15		

To investigate the effects of cross-lingual transfer, we compared classifier performance against the ideal monolingual version: the accuracy which would be achieved if we were able to train and test on target language data. Performance in a cross-lingual setting (training on a different source language and testing on the target language) causes some drop in performance below the ideal case, but this varies with language, and particularly with the closeness of source and target: transfer between languages in the same family gives less of a drop (see Table 5).

We then investigated whether cross-lingual transfer can also help improve accuracy of a monolingual classifier (Table 6). Here, we compare the performance when training and testing on the same language ("only target" column), and when that training set is augmented with data from other languages ("all other & target"). While this succeeds in a few cases (see Bulgarian, Serbian), in general it reduces accuracy below the monolingual case assuming target language data is available, by 4% F1-score on average. In Task T3.2, when applying these techniques to the comment filtering task, we therefore follow

Table 6: Classifier performance as macro-averaged F_1 score and classification accuracy (CA) for the three-way sentiment task with multiple cross-lingual transfer using the LASER library: testing on *target* language data when training on all other *source* languages (Robnik-Šikonja et al., 2020).

	All other & Target		Only Target	
Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
Bosnian	0.64	0.59	0.67	0.64
Bulgarian	0.54	0.56	0.50	0.59
Croatian	0.63	0.57	0.73	0.68
English	0.58	0.60	0.62	0.65
German	0.52	0.59	0.53	0.65
Hungarian	0.59	0.61	0.60	0.67
Polish	0.67	0.63	0.70	0.66
Portugal	0.44	0.39	0.52	0.51
Russian	0.66	0.64	0.70	0.70
Serbian	0.52	0.49	0.48	0.54
Slovak	0.64	0.61	0.72	0.72
Slovene	0.54	0.50	0.60	0.60
Swedish	0.63	0.59	0.67	0.65
Average improvement	-0.04	-0.07		

the approach of source-to-target transfer between similar languages, and investigate at what point the use of cross-lingual transfer stops adding benefit once enough target language data is available; see (Pelicon et al., in preparation), included as part of Deliverable D3.3.

This work is described in full in (Robnik-Šikonja et al., 2020), attached here as Appendix D.

5 Context analysis

The work described so far treats texts as independent. With UGC of the form observed in news comments, this a reasonable first assumption, but is not sufficient to capture all aspects of meaning: many texts are highly context-dependent. Opinions, for example, are often expressed (and can only be fully understood) by agreeing or disagreeing with another commenter. Similarly, a short context-dependent message might be interpreted as highly offensive or as entirely innocent, depending on the context in which it appears. Another stream of our work therefore focuses on modelling context and understanding its effects on the tasks of interest here (e.g., author profiling, sentiment, opinion analysis etc.).

5.1 Mutimodal neural networks for general context modelling

Many aspects of context can help in understanding and summarising UGC for news. One is the ongoing thread of comments posted so far; another is the news article being commented on. Both have their own complex structure: the comment feed has a thread structure and a temporal structure; and the article may include text body, images and captions. To be able to take these into account, we need a general model which is capable of learning the relations between comments and other related content that may be in difference modalities and have complex temporal or sequential relations. To this end, we have developed a novel neural network architecture, structured to allow fusion of information from sources in different modalities, with different structures, lengths and timescales.

To date, this work has been developed and tested on a different domain, but in a task related to the author profiling work in Section 3 above: integrating information from text transcripts, audio recordings

and video information to detect whether a speaker in a sequentially transcribed conversation is suffering from depression. Results beat the state of the art (see Table 7), and the approach is general enough to be applied to arbitrary text and non-text data: our NN model learns separate models for the specific modalities, and uses a gating mechanism to learn how best to combine them for overall task performance (see Figure 3). We expect the method to be suitable for cross-lingual transfer due to the use of standard NN approaches (based on LSTMs and word embeddings which could be mapped cross-lingually or replaced with cross-lingual BERT/ELMo models), and applicable to the multimodal information present in news articles paired with sequentially-structured comment threads.

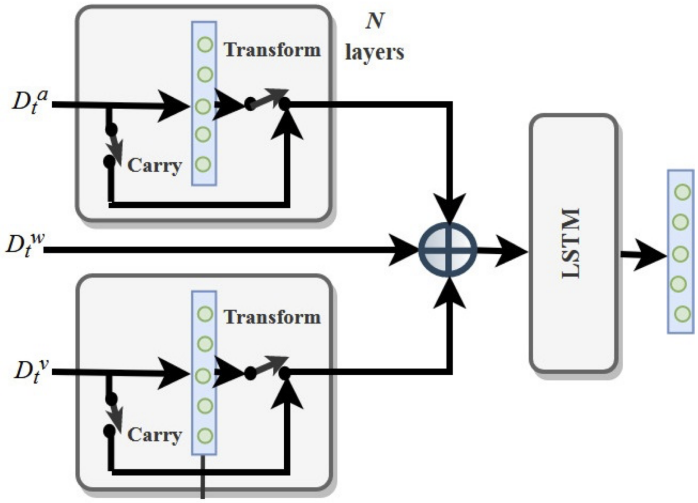

Figure 3: Network structure for multimodal fusion with gating (Rohanian et al., 2019).

Table 7: Classification results, showing the improvements gained by incorporating information from multiple modalities (text, audio, visual) in the new gated model (Rohanian et al., 2019).

Model	Features	F1	Prec.	MAE	RMSE
Baselines					
DAIC Baseline [28]	Audio+Visual	-	-	5.66	7.05
Gong et al. [12]	Text+Audio+Visual	0.60	-	3.96	**4.99**
Alhanai et al. [18]	Text	0.66	0.70	5.09	6.11
Alhanai et al. [18]	Text+Audio	0.75	0.72	5.02	6.04
Williamson et al. [14]	Text	0.67	0.74	3.82	5.06
Williamson et al. [14]	Text+Audio+Visual	0.70	0.78	3.84	5.23
Our Models					
LSTM	Text	0.69	0.68	4.98	6.05
LSTM	Text+Audio	0.67	0.68	5.18	6.40
LSTM	Text+Audio+Visual	0.67	0.63	5.29	6.68
LSTM with Gating	Text+Audio	0.80	0.78	3.66	5.14
LSTM with Gating	Text+Audio+Visual	**0.81**	**0.80**	**3.61**	**4.99**

This work is described in full in (Rohanian et al., 2019), attached here as Appendix E.

5.2 Dialogue structure information

Another more specific line of work we are pursuing is to investigate models of conversation structure: analysing the type of contribution that each comment makes (whether it asks a question, answers a

question, challenges another etc.), and how they relate to each other. This approach is central to much work in dialogue analysis; our first question is whether it is of use in our tasks of interest here.

Our work so far examines the use of dialogue act analysis in author profiling; as in the previous section, our dataset uses spoken dialogue and concerns a specific healthcare profiling task (again, data was processed as part of another project and made available to EMBEDDIA), in this case distinguishing speakers who suffer from dementia. We performed an empirical study of a corpus of conversational transcripts (Pope & Davis, 2011), applying a manual annotation procedure to label utterances with their dialogue act functions, following and slightly adapting a standard coding manual (the Switchboard tagset Jurafsky et al., 1997). This allows us to statistically compare the distributions (see Figure 4): results show that the distributions of dialogue acts differ significantly not only for those that might be expected (e.g., clarification questions, signals of non-understanding) but in some much more general categories (e.g., types of questions and question-answer sequences). These differences give good discriminative information, allowing us to detect dementia-suffering speakers from controls with 70-80% accuracy (depending on the exact feature set used).

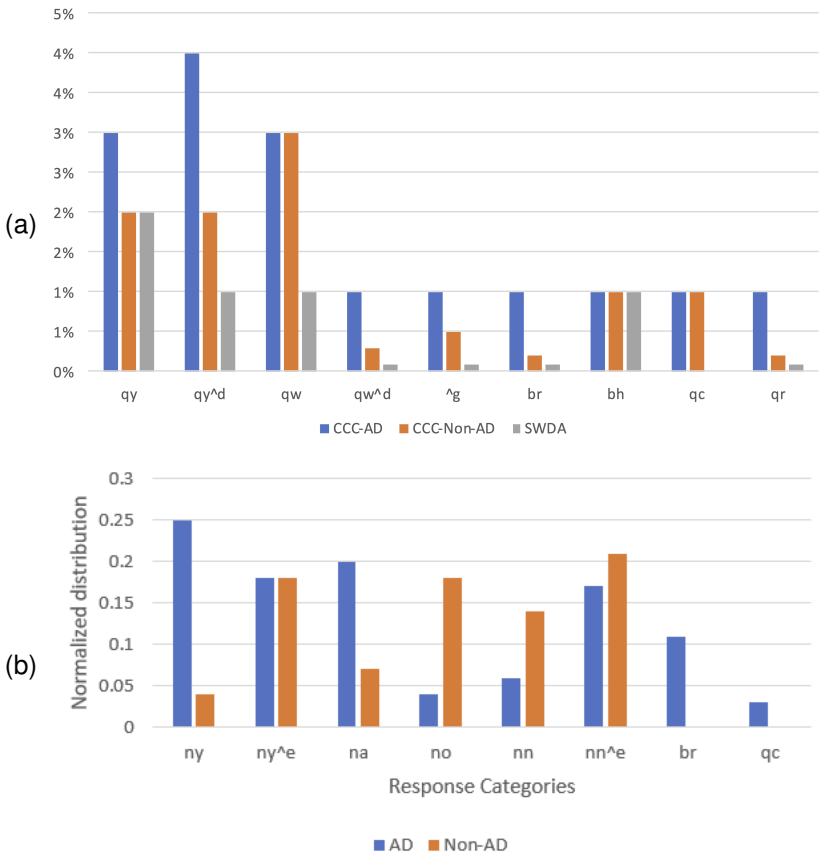

Figure 4: Dialogue act distributions for different speaker and dialogue types (Nasreen et al., 2019a). (a) shows question types, where 'qy' represents yes/no questions, 'qw' wh-questions, '^d' indicates declarative forms; (b) shows answer types in response to yes-no questions, where 'ny' represents positive answers, 'nn' negative answers, 'no' other answers, and '^e' indicates expansions. 'AD' indicates speakers suffering from Alzheimer's Disease; 'Non-AD' controls; 'CCC' is the Carolinas Conversation Corpus being analysed (Pope & Davis, 2011), and 'SWDA' a comparison to a standard two-party telephone conversation corpus (Jurafsky et al., 1997).

This work is described in full in (Nasreen et al., 2019a) and (Nasreen et al., 2019b), attached here as Appendices F and G.

5.3 Comment thread analysis

Our current focus in context analysis is on developing methods for automatically inferring thread structure, i.e. we aim to detect which comment is being responded to at any point. We combine this with the opinion/sentiment methods from Section 4.2 to enrich it with information about stance: whether the response is a positive one (of agreement with or support for the previous comment(er)'s stance) or a negative one (of disagreement).

Automatic thread creation A thread structure is crucial for many of the tasks: understanding which of the previous comments, if any, is the antecedent (parent) comment being responded to. However, in most relevant datasets, including our EMBEDDIA news comment datasets from our news media partners (see Deliverable D3.1), only one level of parent information is annotated. That is, all comments explicitly intended as responses/replies are marked as being associated only with the comment at the start (root) of the thread. (In the case of our news partner datasets, this is all that is captured and stored in their database). However, this does not capture the intentions behind the comments: they are more frequently responding to one of the more recent comments in the same thread. Our initial problem is therefore how to infer the actual response-antecedent relations automatically.

We formulate this as a binary classification problem. That is, given a pair of two comments, our model has to predict whether one is acting as a reply to the other or not. To build a suitable dataset, we use data from the One Million Posts corpus (Schabus et al., 2017) of user comments from an Austrian newspaper website in German, together with EMBEDDIA data from 24sata in Croatian. We treat the first reply comment in any thread as a reply to the parent comment. We then sample negative examples using any other comment from a different thread, for both the One Million Posts and 24sata corpora. To classify the reply-to relation, we use the LASER encoder to produce cross-lingual comment representations (Artetxe & Schwenk, 2019) (shown to be an effective cross-lingual representation and effective in our sentiment analysis experiments in Section 4.2 above). Using this representation, we train a multi-layer perceptron on our positive and negative sampled examples, and achieve 0.63 F_1 score on a test set sampled from both German and Croatian datasets.

Agreement and disagreement detection 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.

We treat stance classification as a two-class classification task: given two comments, the model must predict whether these two comments agree or disagree with each other. To represent a comment, we again use the LASER encoder; the encodings of the two comments are concatenated and passed through a multi-layer perceptron (MLP) to perform the classification. We used two publicly available datasets: the CoREa corpus (Celli et al., 2014) and the YNACC corpus (Napoles et al., 2017), which provide annotation of agreement-disagreement on user comments in Italian and English news respectively. The dataset details are shown in Table 8. We achieve an F_1 score on the test set of 0.80. This shows that automatic agreement-disagreement is possible; we will next turn to more complex models as we expect that performance can be further improved. We are now annotating Croatian data from 24sata (access via partner TRI) for both this task and to give improved information for the thread structure detection task above, and will therefore experiment on this data when available.

Source	Train		Val		Test	
	#comments	#comment-pairs	#comments	#comment-pairs	#comments	#comment-pairs
YNACC (EN)	7262	11212	525	783	525	773
CoREa (IT)	784	1419	201	176	205	179
Total	8046	12631	726	959	730	952

Table 8: Data distribution of existing resources for news comments.

6 Associated outputs

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

Description	URL	Availability
Code for author profiling (PAN 2019)	github.com/EMBEDDIA/PAN2019	Public (MIT)
Code for author profiling (CLIN 2019)	github.com/EMBEDDIA/CLIN29	Public (MIT)
Code for opinion detection	github.com/EMBEDDIA/opinion-detection	To become public*
Code for thread reconstruction	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
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 (CLIN9).	Published	Appendix A
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).	Published	Appendix B
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).	Published	Appendix C
Concannon, S., & Purver, M. (in preparation). Detecting patient opinion in social media [EXACT TITLE ANONYMISED FOR REVIEW]. Draft, for submission to BMJ Quality & Safety.	Draft	(available on request)
Robnik-Šikonja, M., Reba, K., & Mozetić, I. (2020). Cross-lingual transfer of Twitter sentiment models using a common vector space. Submitted.	Submitted	Appendix D
Rohanian, M., Hough, J., & Purver, M. (2019, September). Detecting depression with word-level multimodal fusion. In Proceedings of INTERSPEECH (pp. 1443–1447). Graz, Austria: ISCA.	Published	Appendix E
Nasreen, S., Purver, M., & Hough, J. (2019a, September). 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. London, United Kingdom: SEMDIAL.	Published	Appendix F
Nasreen, S., Purver, M., & Hough, J. (2019b, October). Interaction patterns in conversations with Alzheimer's patients. Abstract and poster, presented at the 7th International Conference on Statistical Language and Speech Processing. Ljubljana, Slovenia.	Presented	Appendix G

7 Conclusions and further work

The objective of this task 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 and 4 show, we have succeeded in developing classifiers for a range of suitable analysis tasks: author type, sentiment, opinion and stance analysis. These all use methods that can be combined with WP1's results on cross-lingual embeddings to produce cross-lingual versions for transfer to EMBEDDIA project languages. Section 4.2 shows that this can indeed be achieved, with little drop in accuracy, for the task of sentiment analysis. We have also developed methods for incorporation of contextual information into the classifier models, shown that such context helps in author profiling, and begun work on inferring context structure to help stance detection (Section 5).

Next steps will extend the work on context modelling to improve our models for inter-personal and inter-comment stance analysis, in order to provide richer information to Task T3.3. We will also extend our work on topic detection, which has so far been limited to using conventional non-embedding-based methods (see Section 4.1), by combining with work in WP4, and applying to our news domain and multilingual setting. We will incorporate further advances from WP1, WP2 and WP4 to improve performance, and transfer the results of this task into the more end-user-centered implementations under development in Tasks T3.2 and T3.3; these in turn will form components of the Media Assistant in WP6.

Bibliography

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, May). 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.

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, December). 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)

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).

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.

Greaves, F., Laverty, A., Ramirez Cano, D., Moilanen, K., Pulman, S., Darzi1, A., & Millett, C. (2014). Tweets about hospital quality: A mixed methods study. *BMJ Quality & Safety*, *23*, 838-846.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... Greaves, F. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Quality & Safety*, *25*, 404-413.

Hovy, D., & Søgaard, A. (2015, July). 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.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual.

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

Li, X., Bing, L., Lam, W., & Shi, B. (2018, July). 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.

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.

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).*

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.

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

Müller, S., Huonder, T., Deriu, J., & Cieliebak, M. (2017, August). 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, September). 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, October). 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., Pollak, S., & Purver, M. (in preparation). *Zero-shot cross-lingual content filtering: Offensive language and hate speech detection.* (Draft)

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

Pizarro, J. (2019). Using n-grams to detect bots on Twitter. In *Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019).*

Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, *7*(1), 143–161.

Purver, M., & Battersby, S. (2012, April). 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., 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.*

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

Rosenthal, S., Farra, N., & Nakov, P. (2017, August). 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.

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).

Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., & Fernández, R. (2019, June). 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.

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).*

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).

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2016, December). 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.

Appendix A: Pooled LSTM for Dutch cross-genre gender classification

Pooled LSTM for Dutch cross-genre gender classification

Matej Martinc[1] and Senja Pollak[1,2]
matej.martinc@ijs.si, senja.pollak@ijs.si

[1] Jožef Stefan Institute,Jamova 39, 1000 Ljubljana, Slovenia
[2] Usher Institute, Medical school, University of Edinburgh, UK

Abstract. We present the results of cross-genre and in-genre gender classification performed on the data sets of Dutch tweets, YouTube comments and news prepared for the CLIN 2019 shared task. We propose a recurrent neural network architecture for gender classification, in which the input word and part-of-speech sequences are fed to the LSTM layer, which is followed by average and max pooling layers. The best cross-genre accuracy of 55.2% was achieved by the model trained on YouTube comments and tweets, and tested on the balanced news corpus, while the best in-genre accuracy of 61.33% was achieved on YouTube comments. Overall, the proposed approach ranked 2nd in the global cross-genre ranking and 6th in the global in-genre ranking of CLIN 2019 shared task.

1 Introduction

Author profiling (AP) is a well-established subfield of natural language processing with a thriving community gathering data, organizing shared tasks and publishing about this topic. AP entails the prediction of an author's profile - i.e. demographic and/or psychological characteristics of the author - based on the text that he/she has written. The single most prominent author profiling task is gender classification, other tasks include the prediction of age, personality, region of origin and mental health of an author.

Gender prediction became a mainstream research topic with the influential work by Koppel et al. (2002). Based on the experiments on a subset of the British National Corpus, they found that women have a more relational writing style (e.g., using more pronouns) and men have a more informational writing style (e.g., using more determiners). Later gender prediction research remained focused on English, but in the last few years, more languages have received attention in the context of author profiling (Rangel et al., 2015, 2016), with the publication of the TwiSty corpus containing gender information on Twitter authors for six languages (Verhoeven et al., 2016) as a highlight so far.

A recent study by van der Goot et al. (2018) calls the cross-genre transferability of machine learning approaches to gender prediction into question by noticing that most of these approaches has typically focused on lexical and specialized social network features, which boosted the performance of the approaches, but on the other hand also made the approaches highly genre and topic dependent. To solve this problem, a fairly new development in the field of AP is the search for data set independent features and approaches, capable of capturing the most generic differences between male and female writing, which transfer well across different genres and languages (Dell Orletta and Nissim, 2018). This is also the main focus of the present research, in which we primarily deal with the development and testing of the system for Dutch cross-genre gender classification. In contrast to the majority of the best performing systems in the field of AP, which use hand-crafted features and traditional classifiers such as Support vector machines (SVM) and Logistic regression (Rangel et al., 2017), we opted for the neural classifier and automated feature engineering.

This paper is structured as follows. The findings from the related work are presented in Section 2. The data sets and the methodology used are presented in Section 3. Results are presented in Section 4, while in Section 5 we conclude the paper and present plans for the future work.

2 Related work

The lively AP community is centered around a series of scientific events and shared tasks on digital text forensic, such as PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[1] and VarDial (Varieties and Dialects)[2] (Zampieri et al., 2014). While VarDial is more focused on the identification of language varieties and dialects, most past PAN AP shared tasks were centered around gender classification.

The first PAN event took place in 2011 and the first AP shared task was organized in 2013 (Rangel et al., 2013). From the beginning, the PAN shared task was multilingual (Rangel et al., 2013, 2014, 2015, 2016, 2017, 2018) and two of the past competitions also had a cross-genre setting (Rangel et al., 2014, 2016). Another shared task dedicated to cross-genre gender classification on Italian documents was the EVALITA 2018 cross-genre gender prediction (GxG) task (Dell Orletta and Nissim, 2018).

The most popular approach to gender classification usually relies on bag-of-words features and SVM classifiers. For instance, winners of the PAN 2017 competition (Basile et al., 2017) used an SVM based system with very simple features (just word unigrams, bigrams and character three- to five-grams).

Some quite successful attempts of tackling the gender classification with neural networks have also been reported. A system consisting of a recurrent neural network (RNN) layer, a convolutional neural network (CNN) layer, and an attention mechanism proposed by Miura et al. (2017) ranked fourth in the PAN

[1] http://pan.webis.de/
[2] http://corporavm.uni-koeln.de/vardial/sharedtask.html

2017 shared task. In the PAN 2018 multimodal gender classification task (Rangel et al., 2018), where the task was to predict the gender of the Twitter user from their tweets and published images, deep learning approaches were prevailing and the overall winners used RNN for texts and CNN for images (Takahashi et al., 2018).

Another related research we looked at was the use of part-of-speech (POS) tags in existing gender classification approaches, since we hypothesized that POS based features would be less topic and genre-dependent, and therefore appropriate for the cross-genre task at hand. Mukherjee and Liu (2010) showed that sequences of POS tags can be successfully used for gender prediction as a standalone feature or in combination with other features. POS tag sequences were also successfully used in combination with other features in the PAN 2017 AP shared task by Martinc et al. (2017), who overall ranked second in the competition and also tested their model in a cross-genre setting (Martinc and Pollak, 2018).

3 Experimental setup

This section describes the data sets, methodology and the conducted experiments.

3.1 Data sets

CLIN 2018 shared task organizers provided six data sets from three different genres. Altogether, they provided 30,000 tweets, 19,658 YouTube comments and 2,832 news, each of them split into a gender labeled train set and an unlabeled test set. All data sets are balanced in terms of number of documents written by male and female authors. A more detailed description of all the data sets in terms of document size and word length is given in Table 1.

Dataset	Documents	Words
Twitter train	20,000	380,074
Twitter test	10,000	192,306
YouTube train	14,744	280,498
YouTube test	4,914	87,038
News train	1,832	336,602
News test	1,000	401,235

Table 1. Data sets used in the experiments

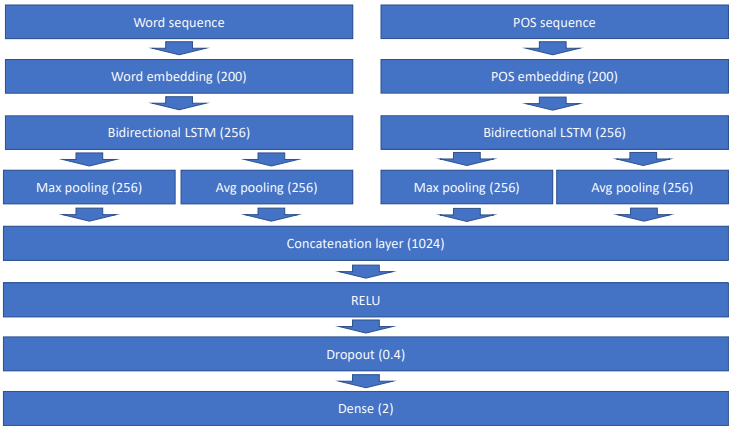

Fig. 1. Infrastructure of the proposed Pooled LSTM network

3.2 Methodology

Altogether, six classification models, three in-genre and three cross-genre, were trained and later used for prediction in our experiments. For the in-genre experiments, the train set for a specific genre was randomly split into a train set containing 90% of the documents and a validation set containing 10% of the documents. For the cross-genre experiments, we trained the Twitter cross-genre model on a concatenation of YouTube and news train sets (Twitter train set was used as a validation set during training), YouTube cross-genre model was trained on a concatenation of Twitter and news train sets (YouTube train set was used as a validation set during training) and news cross-genre model was trained on tweets and YouTube comments (news train set was used as a validation set during training).

Text preprocessing is light, we only replace hashtags in some of the data sets with *#HASHTAG* tokens, URLs with *HTTPURL* tokens and mentions with *@MENTION* tokens. We also limit the text vocabulary to 30,000 most frequent words and replace the rest with the <*unk*> token.

We decided on a neural approach to the task at hand, mostly because of the relatively large sizes of the available train and test sets (described in Section 3.1). Taking into the consideration some of the findings from the related work, we opted for the bidirectional recurrent architecture, which was successfully employed for gender prediction in the past (Miura et al., 2017; Takahashi et al., 2018). Initial experiments and previous research (Martinc et al., 2017; Martinc and Pollak, 2018) also suggested that adding POS tag information improves the performance of the model (especially in the cross-genre setting), therefore POS sequences are fed to the network together with the preprocessed texts.

	Twitter	YouTube	News	Average
Validation set in-genre	0.6245	0.6270	**0.6477**	0.6331
Validation set cross-genre	0.5473	**0.5580**	0.5573	0.5542
Official test set in-genre	0.6099	**0.6133**	0.5990	0.6074
Official test set cross-genre	0.5427	0.5507	**0.5520**	0.5485

Table 2. Results of the in-genre and cross-genre classification

Embedding vectors of size 200 are produced for input word and POS tag sequences, with the help of two randomly initialized embedding layers, and then fed to two distinct Bidirectional Long short-term memory networks (BiLSTM) with 256 neurons, which both produce a two dimensional matrix (with the time-step dimension and the feature vector dimension) representation for every token in the sequence. In order to find the words/POS tags with the highest predictive power, we use an approach similar to the one proposed by Lai et al. (2015), and employ one-dimensional max pooling and average-pooling operations (Collobert et al., 2011) on the time-step dimension to obtain two fixed-length vectors for each of the inputs.

The four resulting vectors are concatenated and fed into the rectified linear unit (RELU) activation function, on the output of which we conduct a dropout operation, in which 40% of input units are dropped in order to reduce overfitting. The resulting vector is passed on to a fully connected layer (*Dense*) responsible for producing the final binary gender prediction.

We use the Python Pytorch library (Paszke et al., 2017) for the implementation of the system. For optimization, we use an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001. Each of the models is trained on the train set for one hundred epochs and tested on the validation set after every epoch. The model with the best performance on the validation set is chosen for the test set predictions. For POS tagging, a Perceptron tagger from NLTK (Bird and Loper, 2004) is used and for measuring the performance of the classifier, accuracy is used.

4 Results

Classification results are presented in Table 2. On the official test sets, the highest cross-genre accuracy (55.20%) was achieved on news. Slightly worse was the accuracy on the data set of YouTube comments (55.07%), while the accuracy on the tweet test set was almost 1% lower. When it comes to the official in-genre results, the highest accuracy was achieved on the test set of YouTube comments (61.33%) and lowest on news (59.99%).

Results on the validation sets are in all cases better than the results on the official test sets, when same genres and same types of classification on validation and test sets are compared. This suggests some overfitting, which is generally more alarming in the in-genre setting, where the training sets were smaller.

Overfitting is the worst in the news in-genre setting, where the difference in performance on the official test set and validation set is almost 5%.

When we compare these results to the results of other teams in the CLIN shared task, our approach yields good performance in the cross-genre part of the competition, where we ranked second as a team, although it should be mentioned that the first ranked team submitted two runs which both performed better than our submitted run. On the other hand, our approach yields worse results in the in-genre setting, where we ranked sixth out of eight teams with the ninth best run.

5 Conclusion

In this paper we presented the results of the CLIN 2019 cross-genre and in-genre gender classification shared task performed on the data set of Dutch tweets, YouTube comments and news. A neural network architecture, which takes word and POS sequences as input, is capable of detecting relatively good features by performing max and average pooling on the output matrix of the LSTM layer. On the official CLIN 2019 test sets, our team ranked second in the cross-genre setting and sixth in the in-genre setting.

Not surprisingly, the models trained and tested on the same genre achieve much better performance than the models with train and test sets from different genres, even though the train sets in the cross-genre setting are much larger in all the cases. The performance of our classifier is quite consistent across all genres, which is against our expectations, since we expected better performance on the news data set because of the on average much longer documents and therefore more per-instance information for the classifier.

Dutch gender classification is still a tough problem, which becomes clear, if we compare the low performances of all the approaches in the shared task with the performances usually achieved on the English data sets in PAN shared tasks. In order to narrow this gap, for the short term future work we plan to test our approach on other languages, just to get the better picture of the difficulty of cross-genre and in-genre gender classification across different languages. We will also be conducting a comprehensive error analysis, which will help us identify language- and genre-independent features that work well across different genres and languages. In the long term, we will try to improve our approach by testing numerous state-of-the-art neural architectures and employ transfer learning techniques.

Acknowledgments

The work presented in this paper has been supported by European Unions Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial

support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The Titan Xp used for this research was donated by the NVIDIA Corporation.

Bibliography

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Felice Dell Orletta and Malvina Nissim. 2018. Overview of the evalita 2018 cross-genre gender prediction (gxg) task. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy. CEUR. org*.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.

Matej Martinc and Senja Pollak. 2018. Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. 2017. Pan 2017: Author profiling-gender and language variety prediction. *Cappellato et al.[13]*.

Yasuhide Miura, Tomoki Taniguchi, Motoki Taniguchi, and Tomoko Ohkuma. 2017. Author profiling with word+ character neural attention network. In *CLEF (Working Notes)*.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. Available at `https://pytorch.org/`.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Working Notes*. CEUR.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*. CEUR.

Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giancomo Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF 2013 Evaluation Labs and Workshop Working Notes Papers*. CEUR.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *CLEF 2016 Working Notes*. CEUR-WS.org.

Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2018. Text and image synergy with feature cross technique for gender identification. *Working Notes Papers of the CLEF*.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: a multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. ELRA, Portorož, Slovenia.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.

Appendix B: Who Is Hot and Who Is Not? Profiling Celebs on Twitter

Who Is Hot and Who Is Not? Profiling Celebs on Twitter

Notebook for PAN at CLEF 2019

Matej Martinc[1,2], Blaž Škrlj[1,2], and Senja Pollak[1,3]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] Usher Institute of Population Health Sciences, Medical School, University of Edinburgh, UK
matej.martinc@ijs.si, blaz.skrlj@ijs.si, senja.pollak@ijs.si

Abstract We describe the system developed for the Celebrity profiling shared task of PAN 2019, capable of determining the gender, birthyear, occupation and fame of celebrities given their tweets. Our approach is based on a Logistic regression classifier and simple n-gram features. The best performance is achieved on the task of gender prediction, while predicting fame and occupation are slightly harder for the system. The worst performance is unsurprisingly achieved on the task of predicting birthyear, the hardest classification problem with seventy unbalanced classes. The proposed system was 3rd in the global ranking of PAN 2019 Celebrity profiling shared task.

1 Introduction

Author profiling (AP) is a field that deals with learning about the demographics and psychological characteristics of a person based on the text she or he produced. The most common tasks from the field include gender, age and language variety prediction but due to a large quantity of content available from social networks, the number of tasks is growing rapidly.

Most AP research is centered around a series of scientific events and shared tasks on digital text forensics, most popular being the series of scientific events and shared tasks called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[4]. The first PAN event took place in 2011 and the first AP shared task was organized in 2013 [12]. One of the most commonly addressed tasks in PAN is the prediction of an author's gender, although previous shared tasks also included tasks such as age, language variety and personality prediction [11,13]. This year, due to the availability of a new celebrity corpus [18], the number of attributes to predict has increased, and the task includes gender, age, fame and occupation prediction.

[4] http://pan.webis.de/

This paper describes our approach to the Celebrity profiling shared task of PAN 2019 [19], which involves the construction of four classification models for four distinct profiling traits on the celebrity corpus.

The rest of the paper is structured as follows: in Section 2 the findings from the related work are presented. Section 3 describes the corpus and how it was preprocessed. In Section 4 we present the feature engineering and classification methodology, while Section 5 presents the results. After a short Discussion (Section 6), we conclude the paper and present ideas for future work in Section 7.

2 Related Work

The first and most popular task addressed in the field of AP was gender prediction. It became a mainstream research topic with the work by Koppel et al. [5], 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 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. For example, the winners of the PAN 2017 competition [2] used a Support vector machine (SVM) based system with simple features (word unigrams, bigrams and character three- to five-grams). Second ranked team [6] used a Logistic regression classifier and a somewhat more complicated combination of word, character and part-of-speech (POS) n-grams, sentiment from emojis, and character flooding as features. 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 tetragrams features.

PAN 2016 AP shared task also dealt with age classification. The winners in this task [17] used a linear SVM model and employed a variety of features: word, character and POS n-grams, capitalization (of words and sentences), punctuation (final and per sentence), word and text length, vocabulary richness, hapax legomena, emoticons and topic-related words. On the other hand, none of the previous PAN tasks included prediction of fame and occupation. While we are not aware of any study which dealt with the celebrity fame prediction, we acknowledge 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.

3 Dataset Description and Preprocessing

The training set for the PAN 2019 Celebrity profiling shared task consists of English tweets from 33,836 celebrities and contains labels for fame, gender, occupation and birthyear (details of a dataset structure are presented in Table 1 and Figure 1). The number of tweets per author is not constant and all classes are inbalanced. The label

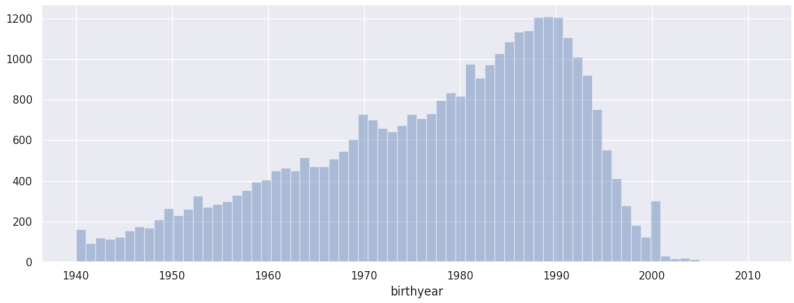

Figure 1. Birthyear distribution in the celebrity corpus

with the most classes is birthyear with 70 distinct values, occupation has 8 classes, while both fame and gender have 3 classes.

Table 1. Fame, gender and occupation distribution of the celebrity corpus

Fame	Gender	Occupation
superstar (7,116)	male (24,221)	sports (13,481)
star (25,230)	female (9.683)	performer (9,899)
rising (1,490)	non-binary (32)	creator (5,475)
/	/	politics (2,835)
/	/	science (818)
/	/	professional (525)
/	/	manager (768)
/	/	religious (35)

First, tweets belonging to the same celebrity are concatenated and used as one document in further processing. If an author has published more than 100 tweets, only first 100 tweets are used, since we believe this is a sufficient amount of content needed for successful profiling of the author and since this procedure drastically decreases the time and space complexity. After that, we employ three distinct preprocessing techniques on the resulting documents, producing three levels of preprocessed texts, which are all used in the feature engineering step:

- *Cleaned level*: replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively.
- *No punctuation level*: removing punctuation from the cleaned level;
- *No stopwords level*: stopwords are removed from the no punctuation level.

4 Feature Construction and Classification Model

Due to findings from the related work (see Section 2), which suggest that reliance on n-gram features and traditional classifiers is still the best approach for most author profiling tasks, we opted for the simplification of the approach we used in the PAN 2017 AP shared task [6]. According to the winners of the PAN 2017 competition [2], adding too sophisticated features negatively affects the performance of the author profiling classification model, therefore our model only contains three different types of n-gram features, which were normalized with the MinMaxScaler from the Scikit-learn library [9]:

- *word unigrams*: calculated on lower-cased No stopwords level, TF-IDF weighting (parameters: minimum document frequency = 10, maximum document frequency = 80%);
- *word bound character tetragrams*: calculated on lower-cased Cleaned level, TF-IDF weighting (parameters: minimum document frequency = 4, maximum document frequency = 80%);
- *suffix character tetragrams* (the last four letters of every word that is at least four characters long [14]): calculated on lower-cased Cleaned level, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%).

We tested several classifiers from Scikit-learn [9]:

- Linear SVM
- SVM with RBF kernel
- Logistic regression
- Random forest
- Gradient boosting

An extensive grid search was performed in order to find the best hyper-parameter configuration for all tested classifiers and the best performing classifier was a Logistic regression with C=1e2 and fit_intercept= False parameters, same as in [6]. The Scikit-learn FeatureUnion[5] class was used to define prior weights for different types of features we used. The weights were adjusted with the help of the following procedure already described in [6]:

1. Initialize all feature weights to 1.0.
2. Iterate the list of features. For every feature repeat adding or subtracting 0.1 to the weight until the accuracy on the validation set is improving. When the best weight is found, move to the next feature on the list.
3. Repeat step 2 until the accuracy cannot be improved anymore.

The weights in our final Logistic regression model were the following:

- word unigrams and word bound character tetragrams: 0.8
- suffix character tetragrams: 0.4

[5] http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

Table 2. Results on the unofficial validation set in terms of cRank (column All) and F1-score (all other columns)

Fame	Gender	Occupation	Birthyear	All
0.7837	0.9017	0.7578	0.0649	0.2092

5 Results

For the unofficial evaluation of our approach, the dataset was randomly split into train (containing 30,000 celebrities) and validation (containing 3,836 celebrities) sets. A separate classification model was trained for each of the classes and we measured performance of the model for each of the classes in terms of weighted F1-score. A measure used for the overall evaluation was cRank, which is a harmonic mean of models performance on each class, or formally:

$$cRank = \frac{4}{\frac{1}{F1_{fame}} + \frac{1}{F1_{gender}} + \frac{1}{F1_{occupation}} + \frac{1}{F1_{birthyear}}},$$

No lenience interval (as is the case in the official PAN 2019 Celebrity profiling shared task evaluation) was used for the birthyear F1-score calculation, therefore prediction was considered incorrect if the exact birthyear was not predicted. Results of the experiments for selected, best performing setting described in Section 4 on the unofficial validation sets in terms of F1-score and cRank are presented in Table 2.

Best results were achieved for the gender prediction task (F1-score of 90.17%), while the hardest attribute to predict was birthyear with an F1-score of only 6.46%. This is not surprising, due to a hard problem of classifying into 70 distinct unbalanced classes. Fame classification appears to be slightly easier for the classifier (F1-score of 78.37%) than the occupation prediction (F1-score of 75.78%) even though the occupation label has eight classes and fame only three. The overall cRank score is low (0.2092) due to the bad performance of the classifier on the task of birthyear prediction.

On the two official test sets the results are very different then on our unofficial validation sets (see table 3). F1-scores for fame, gender and occupation are about 30 percentage points lower on both official test sets, which suggests some serious overfitting. On the other hand, birthyear results on the two official test sets are about 30 percentage points better, most likely due to lenience interval used in the birthyear F1-score calculation, which also positively affected the overall cRank score. All in all, we ranked 3$^{\text{rd}}$ in the official TIRA [10] evaluation.

Table 3. Results on the two official test sets in terms of cRank (column All) and F1-score (all other columns)

	Fame	Gender	Occupation	Birthyear	All
Test dataset 1	0.517	0.580	0.449	0.361	0.462
Test dataset 2	0.507	0.594	0.486	0.347	0.465

6 Discussion

As in last years deep learning is gaining in popularity and achieving state-of-the-art results in a large variety of tasks [16,20,3] and as the celebrity corpus size is relatively large (compared to the PAN 2017 AP datasets), we also considered the neural transfer learning approach BERT (Bidirectional Encoder Representations from Transformers), proposed by [4]. Since the sequence length is limited to 512 characters, we decided to split the text document presenting tweets of each celebrity into chunks equal or smaller than 512 characters and used these chunks as training examples for the classifier. In the prediction phase, the classifier predicted labels for all chunks and majority voting was used to determine the final labels for the entire document. The initial experiments for gender and fame prediction however showed that the BERT classifier is performing much worse (achieving F1-scores of 83.33% and 72.11% for gender and fame, respectively) than the presented Logistic regression classifier. Thus, based on our experiments, we consider that traditional feature engineering techniques are still a better choice for the author profiling on PAN datasets.

7 Conclusion and Future Work

In this paper we have presented our approach to the PAN 2019 Celebrity profiling task, which deals with the prediction of fame, gender, occupation and birthyear for more than 30,000 celebrities. First, we present findings from the related work which suggest that a traditional classification approach with extensive feature engineering presented in this paper is still the preferred approach in the field of AP. We have tested several feature combinations and classifiers and finally selected a Logistic regression classifier with word unigram and character tetragram features, a system very similar to the one we proposed for the PAN 2019 Author profiling task [7].

The Logistic regression classifier and its hyper-parameters were chosen with a grid search but are identical to the study we conducted for the gender classification and language variety shared task in PAN 2017 [6], despite the celebrity corpus being almost ten times bigger than the PAN 2017 author profiling datasets. Because of the large dataset size we also tested the neural transfer learning approach proposed by [4], BERT (Bidirectional Encoder Representations from Transformers). The results were however worse than when the presented Logistic regression classifier was used. Our final results on the two official test sets are F1-scores of 51.7% for fame, 58.0% for gender, 44.9% for occupation and 36.1% for birthyear prediction on the first test dataset, and F1-scores of 50.7% for fame, 59.4% for gender, 48.6% for occupation and 34.7% for birthyear prediction on the second test dataset.

For future work, we believe investigation of potential semantic knowledge's effect on learning, such as explored in [21,15], could also provide valuable insights into parts of the feature space, relevant for learning. We also plan to evaluate the trained gender classification model on other AP datasets with gender labels which do not contain celebrities, in order to determine if the model is transferable.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The work of the second author was funded by the Slovenian Research Agency through a young researcher grant. This paper is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 - project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

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. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
4. 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)
5. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
6. 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)
7. Martinc, M., Škrlj, B., Pollak, S.: Fake or not: Distinguishing between bots, males and females. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019) (2019)
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., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
10. 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 - Lessons Learned from 20 Years of CLEF. Springer (2019)
11. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Working Notes. CEUR (2015)
12. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
13. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)

14. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 93–102 (2015), http://aclweb.org/anthology/N/N15/N15-1010.pdf
15. 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)
16. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1422–1432 (2015)
17. 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)
18. Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. In: Proceedings of ACL 2019 (to appear) (2019)
19. Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
20. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
21. Škrlj, B., Kralj, J., Lavrač, N., Pollak, S.: Towards robust text classification with semantics-aware recurrent neural architecture. Machine Learning and Knowledge Extraction 1(2), 575–589 (2019), http://www.mdpi.com/2504-4990/1/2/34

Appendix C: Fake or Not: Distinguishing Between Bots, Males and Females

Fake or Not: Distinguishing Between Bots, Males and Females

Notebook for PAN at CLEF 2019

Matej Martinc[1,2], Blaž Škrlj[1,2], and Senja Pollak[1,3]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] Usher Institute of Population Health Sciences, Medical School, University of Edinburgh, UK
matej.martinc@ijs.si, blaz.skrlj@ijs.si, senja.pollak@ijs.si

Abstract For the PAN 2019 Author profiling task, we present a two step author profiling system which in the first step distinguishes between bots and humans, and in the second step determines the gender of the human authors. The system relies on a Logistic Regression classifier and employs a number of different word and character n-gram features and a simple type-to-token-ratio feature, which proved useful for the bot prediction task. Experiments show that on the provided datasets of tweets, distinguishing between bots and humans is an easier task than determining the gender of the human authors. The proposed approach was 16th in the global ranking of PAN 2019 Author profiling shared task.

1 Introduction

Social media enables members to interact and share content in an online environment but has recently seen a rise in automated social accounts linked to spamming, fake news dissemination and even manipulation of public opinion. This has had a negative effect on the level of the online discourse and also threatens services such as advertising and search for reliable content [3]. To counteract this tendency, social media companies and the research community have proposed several approaches to identify these bots automatically. This detection relies on differences in content produced by humans and bots and also on differences in an online behaviour.

Once a social media user is successfully identified as human, another field of research, generally known as author profiling (AP), deals with learning about the demographics and psychological characteristics of a person based on the text she or he produced. This type of research has already shown a potential for applications in marketing, social and psychological research, security, and medical diagnosis. The most commonly addressed task in AP is the prediction of an author's gender, which has been the main focus of a series of scientific events and shared tasks on digital text forensic called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[4] since

[4] http://pan.webis.de/

2011, when the first PAN event took place. The first AP shared task was organized in 2013 [19].

In this paper, we describe our approach to the PAN 2019 AP shared task [18] which deals with the construction of a two step prediction model. In the first step, the system distinguishes between bots and humans and in the second step it determines the gender of human Twitter users. The rest of the paper is structured as follows: in Section 2 the findings from related work are presented. Section 3 describes the corpus and how it was preprocessed. In Section 4 we present the methodology, Section 5 presents the results, while in Section 6 we display the results of the conducted ablation study. In Section 7, we conclude the paper and present ideas for a future work.

2 Related Work

A very successful strategy for detecting bots on Twitter was proposed by [9] and is based on the deployment of honeypots for harvesting deceptive spam profiles on social media. Harvested spammers were then analyzed and findings were used in the implementation of classifiers capable of detecting new bot spammers. For classification, they used text features such as n-grams and also meta statistical features, such as the ratio between the number of URLs in the 20 most recently posted tweets and the number of tweets, and the ratio between the number of unique URLs in the 20 most recently posted tweets and the number of tweets. They report the F1-score of 88.80% achieved with the Weka Decorate meta-learner. A more recent classification approach which relied on statistical meta features (age of the account, number of tweets, followers-to-friends ratio, retweets per tweet...) was proposed by [6]. They achieved an accuracy of 86.44% in the 5-fold cross validation setting with a Random Forest classifier.

Another interesting approach was proposed by [5] who among other features (e.g., average number of hashtags and repeated tweets, latent Dirichlet allocation identified topics, graph-theoretic statistics...) also leveraged sentiment-related factors for bot identification. They used a Gradient boosting classifier and also employed statistical features derived from text, such as average number of hashtags, average number of user mentions, links and emoticons.

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. [3] 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 dataset.

Gender prediction became a mainstream research topic with the work by Koppel et al. [7]. Based on experiments on a subset of the British National Corpus, they found that women have a more relational writing style (e.g. using more pronouns) and men have a more informational writing style (e.g. using more determiners). Later gender prediction research remained focused on English, yet the attention quickly shifted to social media applications [2,23,15] and other languages. The most relevant findings for the gender classification task at hand comes from PAN shared tasks in 2016 and 2017 [21,20], where one of the goals was to predict gender of the user on English and

Table 1. PAN 2019 training set statistics

Language	Bots	Male humans	Female humans	All authors	All tweets
English	2,060	1,030	1,030	4,120	412,000
Spanish	1,500	750	750	3,000	300,000

Spanish tweet datasets. In PAN 2016, the best score was achieved by [13], who used word unigrams, word bigrams and character tetragrams features. They used Logistic Regression classifier for learning. A somewhat similar Support vector machine (SVM) based system with simpler features (word unigrams, bigrams and character three- to five-grams) was used by the winners of the PAN 2017 competition [1]. Second ranked team in the PAN 2017 competition [11] also used a combination of word and character n-grams [11], as well as POS n-grams, sentiment from emojis and character flooding as features in the Logistic Regression classifier.

3 Dataset Description and Preprocessing

PAN 2019 training set consists of tweets in English and Spanish languages grouped by tweet authors (100 tweets per author) with gender and type labels (Table 1). Gender and type categories are balanced in both languages. We used this training set in our experiments for feature engineering, parameter tuning and training of the classification models.

First, all tweets belonging to the same author are concatenated and used as one document in further processing. After that, three distinct dataset transformations were employed on the documents and all these three levels of preprocessing were used in the feature engineering step:

- *Cleaned level*: replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively.
- *No punctuation level*: removing punctuation from the cleaned level;
- *No stopwords level*: stopwords are removed from the no punctuation level.

4 Feature Construction and Classification Model

Our feature construction and classification approach can be considered a simplification of the approach we used in the PAN 2017 AP shared task [11], since the winners of the PAN 2017 competition [1] conducted experiments which suggest that adding too sophisticated features negatively affects the performance of the gender classification model. For this reason, our model mostly relies on different types of n-grams and the hypothesis was, that the simplification of the model would also improve the performance of the bot classification model.

4.1 Features

The following n-gram features were used in our final model:

- *word unigrams*: calculated on lower-cased no stopwords level, TF-IDF weighting (parameters: minimum document frequency = 10, maximum document frequency = 80%);
- *word bigrams*: calculated on lower-cased no punctuation level, TF-IDF weighting (parameters: minimum document frequency = 20, maximum document frequency = 50%);
- *word bound character tetragrams*: calculated on lower-cased cleaned level, TF-IDF weighting (parameters: minimum document frequency = 4, maximum document frequency = 80%);
- *suffix character tetragrams* (the last four letters of every word that is at least four characters long [22]): calculated on lower-cased Tweets-cleaned, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%).

The only somewhat more sophisticated feature used in the experiments was calculated on the cleaned level and was inserted in order to improve the performance of the bot classification model:

- *Type-to-token ratio*: calculated by dividing the number of distinct words in the document by the number of all words in the document. The intuition behind this feature is that bots tend to have a higher word repetition frequency and limited vocabulary, therefore low type-to-token ratio could be a good indication that text was produced by a non-human.

All features were normalized with the MinMaxScaler from the Scikit-learn library [14]. For example, a vector x was rescaled as:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}; \tag{1}$$

yielding feature in range [0,1] (if feature values are all positive).

4.2 Classification Model

Several classifiers from Scikit-learn and libSVM were tested:

- Linear SVM [4]
- SVM with RBF kernel [4]
- Logistic Regression [14]
- Random Forest [14]
- Gradient boosting [14]

We performed an extensive grid search to find the best hyper-parameter configuration for all tested classifiers. Best results were obtained with the Logistic Regression with C=1e2 and fit_intercept=False parameters. The Scikit-learn FeatureUnion[5] class also allows to define weights for different types of features we used, which influence the penalties given to specific features during the training process. The weights were adjusted with the help of the following procedure already described in [11]:

1. Initialize all feature weights to 1.0.
2. Iterate the list of features. For every feature repeat adding or subtracting 0.1 to the weight until the accuracy on the validation set is improving. When the best weight is found, move to the next feature on the list.
3. Repeat step 2 until the accuracy cannot be improved anymore.

The weights in our final Logistic Regression model were the following:

- word unigrams and word bound character tetragrams: 0.8
- suffix character tetragrams: 0.4
- type-to-token ratio: 0.3
- word bigrams: 0.1

This weight configuration proved optimal for both classification tasks and both languages and is almost identical to the configuration used in [11].

5 Experiments and Results

English and Spanish tweet datasets were split into train (containing 2,880 authors for English and 2,080 authors for Spanish) and validation (containing 1,240 authors for English and 920 authors for Spanish) sets according to the recommendation of the PAN organizers to avoid overfitting. In the training and validation experiments, gender and bot classification are considered as separate problems, while the predictions on the official test sets were generated in a sequential order, by first determining if an author is either a human or a bot and then conducting gender classification for authors identified as humans. Results of the experiments on the unofficial validation sets and official test sets in terms of accuracy are presented in Table 2. Both classes are balanced, so for bot and gender classification the majority classifier's accuracy is 0.50. On the unofficial validation sets, distinguishing between bots and humans is an easier task for the classifier, achieving 90.16% accuracy on English and 88.04% accuracy on Spanish. Accuracies for gender classification are lower with the classifier achieving 79.52% accuracy on English and 66.96% accuracy on Spanish. This difference in accuracy could also be partially contributed to smaller training set sizes for gender classification. The Spanish gender classification results are also much lower than previous results with a very similar classifier achieved in the scope of the PAN 2017 gender profiling task [11].

On the official test sets, the accuracies of English and Spanish bot classification are lower (89.39% and 87.44% respectively), which might suggest some overfitting. On the other hand, gender classification results are better on the official test sets for both

[5] http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

languages. While on the English official gender classification test set the accuracy is marginally better, the difference on Spanish is almost 9 percentage points. All in all, we ranked 16^{th} in the official TIRA [16] evaluation, beating the LDSE [17] and word embeddings baselines but falling behind the character and word n-gram baselines.

6 Ablation Study

In order to evaluate the contribution of type-to-token and n-gram features in both classification tasks, an ablation study was conducted. Table 3 presents results for three feature configurations. While using only the type-to-token ratio feature for classification produces classification accuracies very similar to the majority classifier (see column *No n-grams*), combining this feature with n-gram features on average improves bot classification accuracy by 0.35 percentage point. On the other hand, type-to-token ratio feature negatively affects gender classification accuracy, reducing it on average by 0.34 percentage point.

The largest gains in accuracy, when the type-to-token ratio feature is used, are achieved on the English bot classification task (gain of 0.81 percentage point). On the other hand, on the Spanish bot classification task the type-to-token ratio feature marginally reduces the accuracy (reduction of 0.11 percentage point) of the classifier. When it comes to gender classification, the results of the ablation study show that the type-to-token ratio feature has a marginally positive effect on the Spanish dataset (gain of 0.21 percentage point) but also reduces the accuracy of the English gender classification by about 0.5 percentage point.

7 Conclusion and Future Work

In this paper we have presented our approach to the PAN 2019 AP task, which deals with distinguishing between humans and bots and with determining the gender of the human authors. First we presented findings from the related work that were considered during the planning phase of our research and influenced this research the most. After that, we described the datasets used in our experiments, the preprocessing and feature engineering techniques used, and the classification algorithms employed in our experiments. Finally, we presented the experiments together with results on the unofficial validation sets.

According to our experiments, distinguishing between bots and humans is a somewhat easier task than distinguishing between male and female humans. We also used

Table 2. Accuracy results on the unofficial validation set and the official test set

	Unofficial		Official	
	Bot	Gender	Bot	Gender
English	0.9016	0.7952	0.8939	0.7989
Spanish	0.8804	0.6696	0.8744	0.7572

exactly the same approach for both classification tasks, even though some related work suggested different sets of features for these two tasks. Different types of word and character n-grams proved as the most useful features in both tasks. There is however a difference in effect of the type-to-token ratio feature when it comes to both tasks. While this feature negatively affects the accuracy of the gender classifier, it does improve the accuracy of the bot classifier by 0.35 percentage point.

Another interesting observation is that even though we conducted an extensive grid search to find the best classifier with the best configuration of hyper-parameters, the final choice is identical to the choice of a classifier and hyper-parameters used in our previous study of gender classification [11], despite the additional problem of bot classification. In addition, very similar setting was also selected as best for our approach in the PAN 2019 Celebrity profiling task [12].

We believe an unexploited opportunity is the body of semantic background knowledge, such as for example the word taxonomies. Approaches such as SRNA [24] could be used to investigate, whether such knowledge contributes to learning for the task at hand.

Another line of future work will deal with the evaluation of the model on additional datasets from other social media platforms besides Twitter in order to test how well our model generalizes across different social media content. For gender identification, online workflows have been proposed [10] in the ClowdFlows environment [8] and we plan to expand the set of workflows to also cover bot identification.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The work of the second author was funded by the Slovenian Research Agency through a young researcher grant. This paper is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 - project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

1. 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)

Table 3. Results of the ablation study on the unofficial validation set

	All features		No type-to-token ratio		No n-grams	
	Bot	Gender	Bot	Gender	Bot	Gender
English	0.9016	0.7952	0.8935	0.8000	0.5040	0.4984
Spanish	0.8804	0.6696	0.8815	0.6717	0.5021	0.5000
Average	**0.8910**	0.7324	0.8875	**0.7358**	0.5031	0.4992

2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. Association for Computational Linguistics (2011)
3. 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)
4. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3), 27 (2011)
5. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 620–627. IEEE Press (2014)
6. 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)
7. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
8. Kranjc, J., Podpečan, V., Lavrač, N.: ClowdFlows: A cloud based scientific workflow platform. In: Flach, P.A., Bie, T.D., Cristianini, N. (eds.) Proc. of ECML/PKDD (2). LNCS, vol. 7524, pp. 816–819. Springer (2012)
9. 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)
10. Martinc, M., Pollak, S.: Reusable workflows for gender prediction. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France (may 2018)
11. 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)
12. Martinc, M., Škrlj, B., Pollak, S.: Who is hot and who is not? profiling celebs on twitter. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019) (2019)
13. 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)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
15. Plank, B., Hovy, D.: Personality traits on twitter -or- how to get 1,500 personality tests in a week. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Lisbon, Portugal (2015)
16. 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 - Lessons Learned from 20 Years of CLEF. Springer (2019)
17. 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)

18. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
19. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
20. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)
21. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: CLEF 2016 Working Notes. CEUR-WS.org (2016)
22. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 93–102 (2015), http://aclweb.org/anthology/N/N15/N15-1010.pdf
23. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9) (2013)
24. Škrlj, B., Kralj, J., Lavrač, N., Pollak, S.: Towards robust text classification with semantics-aware recurrent neural architecture. Machine Learning and Knowledge Extraction 1(2), 575–589 (2019), http://www.mdpi.com/2504-4990/1/2/34

Appendix D: Cross-lingual Transfer of Twitter Sentiment Models Using a Common Vector Space

Cross-lingual Transfer of Twitter Sentiment Models Using a Common Vector Space

Marko Robnik-Šikonja*, Kristjan Reba*, Igor Mozetič†

* University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, SI-1000 Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si kr3377@student.uni-lj.si

† Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
igor.mozetic@ijs.si

Abstract

Word embeddings represent words in a numeric space in such a way that semantic relations between words are encoded as distances and directions in the vector space. Cross-lingual word embeddings map words from one language to the vector space of another language, or words from multiple languages to the same vector space where similar words are aligned. Cross-lingual embeddings can be used to transfer machine learning models between languages and thereby compensate 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 using the joint numerical space for many languages as implemented in the LASER library: the transfer of trained models, and expansion of training sets with instances from other languages. Our experiments show that the transfer of models between similar languages is sensible, while dataset expansion did not increase the predictive performance.

1. Introduction

Word embeddings are representations of words in numerical form, as vectors of typically several hundred dimensions. The vectors are used as an input to machine learning models; for complex language processing tasks these are typically deep neural networks. The embedding vectors are obtained from specialized neural network-based embedding algorithms, e.g., word2vec (Mikolov et al., 2013), or fastText (Bojanowski et al., 2017). Modern 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 the optional help from 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 for joint construction of embeddings in all the involved languages (Artetxe and Schwenk, 2019). The third type of approach is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In this work, we focus on the second group of approaches, i.e. a joint sentence representation for many languages (Artetxe and Schwenk, 2019) as implemented in the LASER library and available for 93 languages.

Sentiment annotation is a costly and lengthy operation, with 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 be therefore useful and would increase the ability to study sentiment-related phenomena for many more languages than possible today.

Using a collection of 13 large Twitter sentiment datasets, annotated in the same manner, we study two modes of cross-lingual transfer based on projections of sentences into a common vector space. The first approach transfers trained models from source to target languages, where the model is trained on source language(s), and used for classification in target language(s) - this model transfer is possible because texts in all involved languages are embedded to the common vector space. The second approach expands the training set with instances from other languages, and then all instances are mapped into the common vector space during neural network training. Additionally, we analyse the quality of representations for the Twitter sentiment classification and compare the common vector space for several languages constructed by LASER library, multilingual BERT, and traditional bag-of-words approach.

The paper is divided into four sections. In Section 2, we present background on different types of cross-lingual embeddings: alignment of monolingual embeddings, building a fixed common vector space for several languages, and multilingual contextual 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 analyse transfer of trained models between languages from the same language group and from a different language group, followed by the expansion of datasets with instances from other languages. We end the exper-

imental part with the evaluation of representation spaces, and compare the common vector space with multilingual BERT model. In Section 5 we summarize the results, draw the conclusions, and present ideas for further work.

2. Background

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 the other, 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 approaches 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 involved languages. The third type of approach is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). Multilingual BERT is typically used as a starting model which is fine-tuned for a particular task, without explicitly extracting embedding vectors.

2.1. Alignment of monolingual embeddings

Cross-lingual alignment methods take precomputed word embeddings for each language and align them with an 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 common 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 implementation of the method described in (Artetxe et al., 2018a), named *vecmap*[1], is able to align monolingual embeddings either using 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 minimizes the sum of squared Euclidean distances. The algorithm always converges but can be caught in a local maximum when the initial solution is poor. To overcome this, several methods (stochastic dictionary introduction, frequency-based vocabulary cutoff, etc) are used that help the algorithm to 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 poor 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 pseudo-inverse approach for the given initial dictionary. Then the optimal dictionary for the given embeddings is computed and the procedure is repeated with the new dictionary.

When constructing mappings between embedding spaces, a bilingual dictionary can be helpful as its entries can be used as anchors for the alignment map for supervised and semi-supervised approaches. However, lately researchers have proposed approaches that do not require the use of bilingual dictionary, but rely on adversarial approach (Conneau et al., 2018) or use the frequencies of the words (Artetxe et al., 2018b) in order to find a required transformation. These are called unsupervised approaches.

2.2. Projecting into a common vector space

To construct a common vector space for all involved 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 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[2] (Language-Agnostic SEntence Representations) is a Facebook research project focusing on joint sentence representation for many languages (Artetxe and Schwenk, 2019). Similarly to machine translation architectures, it uses an encoder-decoder architecture. The encoder is trained on a large parallel corpora, 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 a large number of languages, currently the encoder supports 93 different languages. The resulting joint embedding can be transformed back into a sentence using decoder for the specific language. This allows training a classifier on data from just one language and use it on any language supported by LASER.

2.3. Multilingual BERT

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

[1] https://github.com/artetxem/vecmap

[2] https://github.com/facebookresearch/LASER

the task of LM 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 very common words are kept as single tokens, others are split into subwords (e.g., common stems, prefixes, suffixes—if needed down to a 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 of the parameters of BERT and new class specific weights are fine-tuned jointly to maximize the log-probability of the correct labels.

3. Datasets and experimental settings

In this section we present the experimental data used, the implementation details of the used neural prediction models, and the evaluation metrics.

3.1. 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. Authors studied the annotators agreement on the labelled tweets and discovered that for some languages (English, Russian, Slovak) the SVM classifier achieves significantly lower score than the annotators. This hints that for these languages there might be a room for improvement using better classification model or larger training set.

We cleaned the above datasets by removing the duplicated tweets, web links, and hashtags. Due to low quality of sentiment annotations indicated by low self-agreement and low inter-annotator agreement we removed Albanian and Spanish datasets. The characteristics of the remaining 13 datasets are presented in Table 1.

Language	Negative	Neutral	Positive	All
Bosnian	12.868	11.526	13.711	38.105
Bolgarian	15.140	31.214	20.815	67.169
Croatian	21.068	19.039	43.894	84.001
English	26.674	46.972	29.388	103.034
German	20.617	60.061	28.452	109.130
Hungarian	10.770	22.359	35.376	68.505
Polish	67.083	60.486	96.005	223.574
Portuguese	58.592	53.820	44.981	157.393
Russian	34.252	44.044	29.477	107.773
Serbian	24.860	30.700	16.161	71.721
Slovak	18.716	14.917	36.792	70.425
Slovene	38.975	60.679	34.281	133.935
Sweedish	25.319	17.857	15.371	58.547

Table 1: The number of tweets from each of the category and the overall number of instances for individual languages.

3.2. Implementation details

In our experiments, we use two different types of prediction models, BiLSTM neural networks using common vector space embeddings constructed with the LASER library, and multilingual BERT. The multilingual BERT model is case sensitive, pretrained on 104 languages, has 12 transformer layers, and 110 million parameters. We fine-tune only the last layer of the network, using the batch size of 32, and 3 epochs.

The cross-lingual embeddings from LASER library are pre-trained on 93 languages, using BiLSTM networks, and are stored as 1024 dimensional embedding vectors. Our classification models contain the embedding layer, followed by 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 ReLU activation function and Adam optimizer. The fine-tuning uses batch size of 32 and 10 epochs.

3.3. Evaluation metrics

Following Mozetič et al. (2016) we report $\overline{F_1}$ score which takes both positive and negative sentiment into account, and classification accuracy CA. $F_1(c)$ 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 $[0, 1]$ interval, where 1 means perfect classification and 0 completely wrong predictions. We use F_1 score averaged over positive $(+)$ and negative $(-)$ sentiment class:

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

The classification accuracy CA is defined as the ratio of correctly predicted tweets N_c to all the tweets N:

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

4. Experiments and results

Our experimental evaluation focuses of text representation using embeddings into a common vector space with the LASER library. We conducted several experiments reported below: transfer of models between languages from

the same and different language family, expansion of training sets with different amounts of data from other languages, and comparison of the common space embeddings with multilingual BERT.

4.1. Transfer to the same language family

We first test the transfer of prediction models between similar languages from the same language family. The transfer between similar languages is the most likely to be successful. As source and target languages we tried several combinations of Slavic and Germanic languages. We report the results in Table 2.

		Transfer		Both target	
Source	Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
German	English	0.55	0.59	0.62	0.65
Polish	Russian	0.64	0.59	0.70	0.70
Polish	Slovak	0.63	0.59	0.72	0.72
German	Swedish	0.58	0.57	0.67	0.65
German Swedish	English	0.58	0.60	0.62	0.65
Slovene Serbian	Russian	0.53	0.55	0.70	0.70
Slovene Serbian	Slovak	0.59	0.52	0.72	0.72
Average performance gap		0.09	0.11		

Table 2: The transfer of trained models between languages from the same language family using common vector space (left-hand side) and comparison with both training and testing set from the target language (on the right-hand side).

In each experiment, we use the complete dataset(s) of the source language as the training set, and the complete dataset of the target language as the testing set. We compare the results with the training and testing set from the target language, where 70% of the dataset is used for training and 30% for testing. The later results can be taken as an upper bound of what the transfer models could achieve in an ideal condition.

The results from Table 2 show that there is a gap between transfer learning models and native models from 4% to 20% (on average 9.3% for $\overline{F_1}$ and 11.1% for CA). For direct transfer of models without additional target data these results are encouraging.

4.2. Transfer to different language family

We repeat the experiments we did for languages from the same language family on languages from different language families. The transfer is less likely to be successful in this case and we expect a lower performance in these unfavourable conditions.

The results from Table 3 show that there is a gap between transferred models and native models from 4% to 28% (on average 14% for $\overline{F_1}$ and 15.2% for CA). This gap is significant and makes the resulting transferred models less useful in the target languages. Another observation is that the differences between target languages are significant. It seems that the transfer to Slovak is much less successful than to Russian, while English is in between the two.

		Transfer		Both target	
Source	Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
Russian	English	0.52	0.56	0.62	0.65
English	Russian	0.57	0.58	0.70	0.70
English	Slovak	0.46	0.44	0.72	0.72
Polish, Slovene	English	0.58	0.57	0.62	0.65
German, Swedish	Russian	0.61	0.61	0.70	0.70
English, German	Slovak	0.50	0.47	0.72	0.72
Average performance gap		0.14	0.15		

Table 3: The transfer of trained models between languages from different language groups using the common vector space representation (left-hand side), and comparison with both training and testing set from the target language (on the right-hand side).

4.3. Increasing datasets with several languages

We test possible improvements in prediction performance if we increase the training sets with instances from several related and unrelated languages. The training set in each experiment consists of instances from several languages projected into the common vector space and also 70% of the target language dataset. The remaining 30% of target language instances are used as the testing set. As the text representation we use projection into the common vector space computed with the LASER library.

The results from Table 4 show a gap between learning models using the expanded datasets and native models (from 2% to 7%, on average 3% for F_1 and 5.7% for CA). These results indicate that the tested expansion of datasets was unsuccessful, i.e. the provided amount of instances from the target language was already sufficient for successful learning. The additional instances from other languages are likely to be of lower quality then the native instances and therefore decrease the performance.

To test an even larger expansion of the training sets, we trained models on all other languages and 70% of the target language, while testing them on the remaining 30% of the target language. The results are presented in Table 5.

The results show that using many languages and significant enlargement of datasets can be successful. For Bulgarian and Serbian training on many languages gives higher $\overline{F_1}$ score (but not CA) than training only on the target language. For all other languages, the tried expansions of training sets are unsuccessful and the difference to native models is on average 3.5% for $\overline{F_1}$ score and 6.8% for CA.

		Expanded		Only target	
Source	Target	$\overline{F_1}$	CA	$\overline{F_1}$	CA
English, Croatian, Slovene	Slovene	0.58	0.53	0.60	0.60
English, Croatian, Serbian, Hungarian, Slovak	Slovak	0.67	0.65	0.72	0.72
English, Croatian, Russian	Russian	0.67	0.65	0.70	0.70
Russian, Swedish, English	English	0.60	0.61	0.62	0.65
Average improvement		-0.03	-0.06		

Table 4: The expansion of training sets with instances from several languages projected into the common vector space using the LASER library (left-hand side) and comparison with training and testing set from the same language (on the right-hand side).

Target	All other & Target $\overline{F_1}$	CA	Only Target $\overline{F_1}$	CA
Bosnian	0.64	0.59	0.67	0.64
Bulgarian	0.54	0.56	0.50	0.59
Croatian	0.63	0.57	0.73	0.68
English	0.58	0.60	0.62	0.65
German	0.52	0.59	0.53	0.65
Hungarian	0.59	0.61	0.60	0.67
Polish	0.67	0.63	0.70	0.66
Portugal	0.44	0.39	0.52	0.51
Russian	0.66	0.64	0.70	0.70
Serbian	0.52	0.49	0.48	0.54
Slovak	0.64	0.61	0.72	0.72
Slovene	0.54	0.50	0.60	0.60
Swedish	0.63	0.59	0.67	0.65
Average improvement	-0.04	-0.07		

Table 5: The expansion of training sets with instances from all other languages mapped into the common vector space using the LASER library (left-hand side) and comparison with training and testing set from the same language (on the right-hand side).

Language	LASER $\overline{F_1}$	LASER CA	mBERT $\overline{F_1}$	mBERT CA	SVM $\overline{F_1}$	SVM CA
Bosnian	**0.67**	0.64	0.65	**0.66**	0.61	0.56
Bulgarian	0.50	0.59	**0.58**	**0.60**	0.52	0.54
Croatian	**0.73**	**0.68**	0.64	**0.68**	0.61	0.56
English	0.62	0.65	**0.72**	**0.71**	0.63	0.64
German	0.53	0.65	**0.66**	**0.66**	0.54	0.61
Hungarian	0.60	0.67	**0.65**	**0.69**	0.64	0.67
Polish	**0.70**	0.66	**0.70**	**0.73**	0.68	0.63
Portugal	0.52	0.51	**0.66**	**0.67**	0.55	0.51
Russian	0.70	0.70	**0.74**	**0.75**	0.61	0.60
Serbian	0.48	0.54	0.56	0.54	**0.61**	**0.56**
Slovak	**0.72**	0.72	0.70	**0.75**	0.68	0.68
Slovene	0.60	0.60	**0.66**	**0.64**	0.55	0.54
Swedish	**0.67**	0.65	0.64	**0.66**	0.66	0.62
Average	0.62	0.64	**0.66**	**0.67**	0.61	0.59

Table 6: Comparison of different representations: supervised mapping into a common vector space with the LASER library, multilingual BERT, and bag-of-ngrams with SVM classifier. The best score for each language and metric is in bold.

4.4. Comparing embeddings

In our final experiment, we compare embeddings into a common vector space obtained with LASER library with multilingual BERT. Note that in this experiment there is no transfer between different languages but only a test of the quality of the representation, i.e. embeddings. The training set in each experiment consists of randomly chosen 70% of the dataset for each language, while the remaining 30% of instances are used as the testing set. As a baseline, we report the results of the SVM model without neural embeddings that uses Delta TF-IDF weighted bag-of-ngrams representation with substantial preprocessing of tweets (Mozetič et al., 2016). These results are not entirely comparable with our setting as they were obtained with 10-fold stratified blocked cross-validation, while we use a single 70:30 split. Further, the datasets for Bosnian, Croatian, and Serbian language were merged in (Mozetič et al., 2016) due to similarity of these languages, therefore we report the performance on the merged dataset for the SVM classifier. Results are presented in Table 6.

The SVM baseline using bag-of-words representation achieves lower predictive performance than the two neural embedding approaches. We speculate that the main reason is the knowledge about language structure contained in large precomputed 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. The multilingual BERT is the best of the three tested methods, achieving the best average $\overline{F_1}$ and CA scores, as well as the best result in most languages (in bold). The downside is that the fine-tuning and execution of mBERT requires much more computational time compared to precomputed fixed embeddings. Nevertheless, with progress in optimization 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.

5. Conclusions

We studied two approaches to the cross-lingual transfer of Twitter sentiment prediction models based on mappings of words into the common vector space: transfer of trained models, and expansion of datasets with instances from other languages. Our empirical evaluation is based on relatively large datasets of labelled tweets from 13 European languages. As word representations, we used mappings into a common vector space produced by the LASER library. 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, we get on average 9.3% lower $\overline{F_1}$ score and 11.1% lower CA. Using models trained on languages from different language families produces larger differences (on average 14% for $\overline{F_1}$ and 15.2% for 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. Finally, we tested the quality of text representations by comparing cross-lingual joint embedding space of LASER library, multilingual BERT embeddings, and classical bag-of-ngram representation coupled with SVM classifier. The results show that multilingual BERT is the most successful of the three, followed by the common vector space of LASER library, while bag-of-ngrams is almost never competitive. The code of our study is freely available [3].

In future work, we plan to expand the experiments with other embedding techniques, in particular the ELMo con-

[3] https://github.com/kristjanreba/cross-lingual-class

textual embeddings (Peters et al., 2018) together with non-isomorphic cross-lingual transformations that could produce better representations in the joint vector spaces.

Acknowledgements

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103. 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 reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

6. References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of International Conference on Learning Representation ICLR 2018*.

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.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint 1309.4168*.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5).

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

Kristjan Reba. 2019. Cross-lingual classification of tweet sentiment / Medjezikovna klasifikacija sentimenta tvitov. BSc thesis, University of Ljubljana, Faculty of Computer and Information Science. (in Slovene).

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. *Cross-Lingual Word Embeddings*. Morgan & Claypool Publishers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations, ICLR 2019*.

Appendix E: Detecting Depression with Word-Level Multimodal Fusion

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Detecting Depression with Word-Level Multimodal Fusion

Morteza Rohanian, Julian Hough, Matthew Purver

School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{m.rohanian, j.hough, m.purver}@qmul.ac.uk

Abstract

Semi-structured clinical interviews are frequently used diagnostic tools for identifying depression during an assessment phase. In addition to the lexical content of a patient's responses, multimodal cues concurrent with the responses are indicators of their motor and cognitive state, including those derivable from their voice quality and gestural behaviour. In this paper, we use information from different modalities in order to train a classifier capable of detecting the binary state of a subject (clinically depressed or not), as well as the level of their depression. We propose a model that is able to perform modality fusion incrementally after each word in an utterance using a time-dependent recurrent approach in a deep learning set-up. To mitigate noisy modalities, we utilize fusion gates that control the degree to which the audio or visual modality contributes to the final prediction. Our results show the effectiveness of word-level multimodal fusion, achieving state-of-the-art results in depression detection and outperforming early feature-level and late fusion techniques.

Index Terms: depression, recurrent neural networks, computational paralinguistics, modality fusion, gestural behaviour, lexical content

1. Introduction

The automatic diagnosis of depression has gained popularity in recent years: depression has a high degree of public prevalence and is one of the most serious forms of disability worldwide [1]. Diagnosis and assessment for depression is generally based around the judgement of clinicians, and commonly uses semi-structured interviews, guided by predetermined sets of topics, in a clinical set-up.

Depression causes cognitive and motor changes that affect speech production: reduction in verbal activity productivity, prosodic speech irregularities and monotonous speech have all been shown to be symptomatic of depression [2]. Depressed patients' spectral-based features have been observed as changing noticeably in depressive states [3]. Their affective state is also influenced by the condition, indicated through prosodic features [4]. However despite several factors being mildly predictive of a depressive state, it has been claimed that because of the innate differences in speaking manner, no single feature on its own has enough discriminatory power as an indicator of depression [5].

Paralinguistic nonverbal cues have been used as depression markers in clinical sessions. Depressed patients exhibit less facial expressivity [6] and less frequent mouth movement [7]. They are more likely to have impaired attention and keep mutual gaze less frequently [8], turn away their gaze and turn their heads down [6]. In addition to nonverbal behavior, linguistic analysis displays important depression indicators. The lexical content of a patient's utterances in clinical interviews has been shown to be effective in detecting depression [9]. Considering the broad clinical outline of depression, it seems that there are significant benefits to be gained from a *multimodal* approach to detecting depression, integrating features from sets of verbal and nonverbal channels of communication.

2. Previous work on depression and cognitive state detection

Recent experimental work has explored the automatic analysis of depression from multimodal data. There has been work on building systems that classify severity of depression using a wide range of multimodal features. Publicly available multimodal depression datasets, which are collections of clinical interviews, have provided an opportunity to explore a range of experiments on detecting depression. Most current approaches use either *early* feature-level fusion whereby features from the different modalities are combined into a new feature set for classification, or *late* prediction-based fusion whereby separate classifiers are trained on each modality to predict the depression state and the the output of those classifiers are combined into a single prediction. Meng et al. use Partial Least Square (PLS) regression for predicting depression based on each modality and apply a late fusion method for the final prediction [10]. Yu et al. propose a multimodal Hidden Conditional Random Fields (HCRF) model considering question and response pairs [11]. Along the same line, Gong et al. combine topic modeling of question/answer of the interviews with multi-modal text, audio, and video features to predict depression levels [12]. Yang et al. use manually selected features as input into a Deep Convolutional Neural Network (DCNN). The learned features are fed to a Deep Neural Network (DNN) to predict the severity of depression [13].

In terms of the communicative features which aid depression detection, lexical features from the interviewer's utterances are shown to be an informative feature for depression in a multimodal classification task with a staircase Gaussian approach [14]. There has also been work on modelling unimodal sequential input for depression detection. Ma et al. propose an audio based method for depression classification using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for a higher level audio representation [15]. Sun et al. present a unimodal random forest method based on the question/answer characteristics of the interview sessions [16].

Nasir et al. consider the temporal nature of audio/visual modalities using a window-based representation of the features instead of the more common approach of frame-level analysis [17]. Utilizing complementary information from text and audio features, Alhanai et al. proposed a model in which two LSTM branches, one per modality, are integrated via a feed-forward network [18]. However, while this work tries to predict depression based on late or early fusion methods [10, 12] or the sequential nature of their inputs [17, 18], learning the

http://dx.doi.org/10.21437/Interspeech.2019-2285

time-dependent relationships between language, visual and audio features in detecting depression is still unexplored.

In other related tasks using multimodal fusion to predict a cognitive state, there has been work on combining temporal information from two or more modalities in a recurrent approach in audio/visual emotion classification [19] and image captioning tasks [20]. This work demonstrated the ability to learn complicated decision boundaries that other models with different fusion methods have difficulty handling [21]. One major problem these models have is dealing with the different predictive power of each modality and their different levels and types of noise. Adding gating mechanisms has been shown to be effective in dealing with the level of contribution of each modality to the final prediction in different multimodal tasks [22, 23].

Our approach is motivated by some of the recent efforts in multimodal fusion for classifying cognitive states to capture the interaction between modalities in detecting depression and maximise the use and combination of each modality. In this paper we propose a *word-level* multimodal fusion with a simple gating mechanism in a time-dependent recurrent framework, and compare it with early and late fusion techniques.

3. Proposed Approach: Word-level multimodal fusion with gating

To predict the severity of depression based on learning multimodal representations, we explore three techniques for fusion: early, late and a model-based approach in which optimal fusion is learned using a neural network. We explore the use of a gating mechanism to learn how best to filter the visual and auditory modalities' effect on lexical information.

3.1. Pre-processing: Forced Alignment for word timings

An essential part of multimodal representation is to model the inter-modality dynamics: to properly learn the time-dependent interactions between language, visual and audio features and integrate them using timestamps. While in a live system we would use time-stamps from a speech recognizer, for this proof-of-concept study we perform offline forced alignment between text, audio and visual features to get the precise time-stamp of every uttered word. At every time-step, we align words with their matching audio time interval using the Penn Phonetics Lab Forced Aligner (P2FA) [24]. P2FA is a tool that can be applied to align transcriptions to audio files, phoneme by phoneme. Upon manual inspection the forced alignment was performed with high enough accuracy for the fusion study in this paper.

3.2. Gating Mechanism

Data from the three modalities have different effects on the final output and it is important to consider the amount of noise when aggregating them into a representation. Since learned representation for the text can be undermined by corresponding visual and audio modalities, we need to alleviate the effects of noise and overlap during multimodal fusion. One way to overcome this problem is to go beyond naive concatenation of vectors representing either the features themselves, or predictions derived from them, and control the degree to which, the audio and visual data contribute to the final prediction using a simple gating mechanism.

We utilize feed-forward highway layers [25], with gating units which learn to regulate information flow through the network by weighting visual and audio inputs at each time-step.

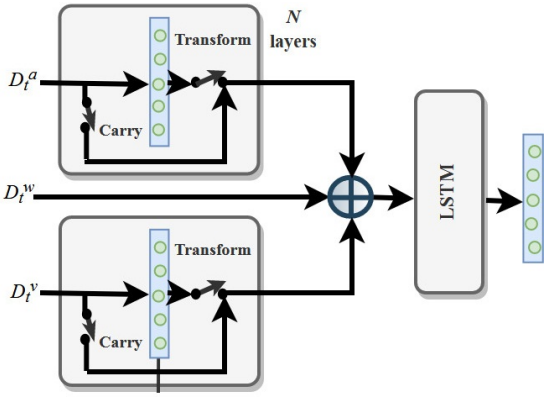

Figure 1: *Word-level multimodal fusion with gating.*

Each highway layer comprises two non-linear transforms: a Carry (Cr) and a Transform (Tr) gate which define the degree to which the output is created by transforming the input and carrying it (how much information should move forward or be changed in successive training epochs). Each layer controls its input vector D_t using the gates and a feed-forward layer H:

$$y = Tr \cdot H + Cr \cdot D_t \quad (1)$$

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 outputs of the transform gates, highway layers can change their performance from layers made of multiple units to layers which only pass their inputs through. As inspired by [25] and to help overcome long-term dependencies earlier in learning, we initialize b_{Tr} with a negative value (biased towards the Carry gate). We use a block of stacked highway layers.

3.3. Model Architecture

We set our model up to learn the most useful interactions between modalities for predicting depression. To achieve this, feature vectors from the three modalities are concatenated to create the input D_t to a word-level LSTM at each time-step t. The overall architecture of our LSTM with Gating model is shown in Figure 1. The gating mechanism is first applied to the audio and visual feature input vectors D_t^a and D_t^v which are passed through N highway layers (where the best value N is determined from optimizing on heldout data) before being concatenated with the current word embedding D_t^w to form the input vector to the LSTM network. After training our LSTM with gating, the resulting Mean Absolute Error (MAE) loss is used as the signal for training our highway layers, employing the REINFORCE rule [26] in a similar way to [27].

Fusion comparison. In addition to testing the effect of using full multi-modality as described compared to combinations of two modalities and single modalities, and also investigating the effect of the gating mechanism, we also compare our model-based fusion technique to two commonly used fusion techniques: early (i.e., feature-based) and late (i.e., decision-based) fusion. In early fusion we integrate features right after extraction (by concatenating them), passing the concatenated

feature vector as input into the LSTM. The late fusion classifier obtains unimodal decision values from three different LSTMs, one for each modality, and then combines their decisions using a weighting mechanism for the final prediction.

4. Experiments

Data. We experiment with datasets from the publicly available Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) with audio, text transcripts and visual features [28]. The DAIC dataset contains clinical interviews, conducted by an animated virtual agent. The training, development, and test sets contain 107, 35, and 47 subjects and the state of the subjects is evaluated based on the PHQ-8 metric [29]. The PHQ-8 assessment rates the severity of symptoms detected in depression, like anxiety, insomnia and agitation to assign a score to a patient based on their level of depression. In addition to binary state of subjects, we predict different degrees of depression at the subject level on the designated test set. The level of depression ranges from 0 to 24 with the range 0-4 regarded as not depressed, 5-9, 10-14 and 15-19 as moderate and 20+ as severe.

4.1. Multimodal features

Lexical Features from Text A Pre-trained GloVe model [30] with a 300-dimensional embedding space was used to extract the lexical feature representations from the transcript. We convert the sequences of responses into word vectors, without considering the queries that led to the responses.[1]

Audio Features A set of audio features are extracted using the COVAREP acoustic analysis framework software [31]. The features include 12 Mel-frequency cepstral coefficients (MFCCs), voiced/unvoiced segmentation and pitch tracking [32], peak slope and maximal dispersion quotients, glottal source parameters (using glottal inverse filtering of GCI synchronous IAIF) [33] and shape parameter of the Liljencrants-Fant model of glottal pulse dynamics [34]. These audio features are extracted based on various attributes of human voice that have been shown to be helpful in detecting depression [5]. Since words are the fundamental units of the input in our models, the interval duration of each word is used as a time interval for capturing these features for each input step. The values for each 10ms frame are averaged to make a single vector for the current word's duration.

Visual Features The visual features are frame-level (20ms window, 10ms shift), provided with the DAIC dataset. They are extracted using the library OpenFace [35] which includes estimates of head position, head rotation, 68 facial landmark locations, gaze tracking, facial action units (FAUs) and HOG features [36]. As with the audio, the average of the frame-level features of the interval duration of each word are used as the visual modality information.

4.2. Implementation and Metrics

All of the experiments are performed without conditioning on speaker identity. The layer sizes and the learning rates are determined using grid search on validation data. The N for Highway networks is an additional hyperparameter required over standard recurrent deep approaches, and 3 was found to be the optimal value. The LSTM models have 128 hidden nodes and are trained using ADAM [37] with learning rate 0.0001. The Mean Absolute Error (MAE) from the ground-truth PHQ-8 assessment scores for each subject is used as the loss function.

For binary classification of depression, we report precision and F1 score and for the PHQ-8 numeric rating accuracy we report the MAE and Root Mean Square Error (RMSE).

4.3. Baseline Models

We compare the performance of our models to the following four models that use the DAIC dataset whose approaches are related to our work: (i) the DAIC baseline with an ensemble of features in a Support Vector Machine (SVM) model which was provided with the dataset [28]; (ii) Gong et al. which uses an ensemble of features with an approach based on topic-modeling [12], (iii) Alhanai et al.'s alternative deep learning model which uses two LSTMs (audio-based and text-based) and a final feed-forward network to model sequences of interactions for detecting depression [18]; (iv) Williamson et al. which performs topic-dependent fusion scoring on text, audio and video [14].

5. Results

Table 1: *Result of the depression classification experiments with our models against state-of-the-art competitors*

Model	Features	F1	Prec.	MAE	RMSE
Baselines					
DAIC Baseline [28]	Audio+Visual	-	-	5.66	7.05
Gong et al. [12]	Text+Audio+Visual	0.60	-	3.96	**4.99**
Alhanai et al. [18]	Text	0.66	0.70	5.09	6.11
Alhanai et al. [18]	Text+Audio	0.75	0.72	5.02	6.04
Williamson et al. [14]	Text	0.67	0.74	3.82	5.06
Williamson et al. [14]	Text+Audio+Visual	0.70	0.78	3.84	5.23
Our Models					
LSTM	Text	0.69	0.68	4.98	6.05
LSTM	Text+Audio	0.67	0.68	5.18	6.40
LSTM	Text+Audio+Visual	0.67	0.63	5.29	6.68
LSTM with Gating	Text+Audio	0.80	0.78	3.66	5.14
LSTM with Gating	Text+Audio+Visual	**0.81**	**0.80**	**3.61**	**4.99**

In Table 1, we present our proposed word-level fusion model's performance against that of baselines and previous state-of-the-art models on depression detection on the provided test set. For detecting depression, our proposed word-level fusion LSTM model with gating achieves an F1 score of 0.81 and MAE of 3.61, outperforming all the baselines. The overall results support our assumption that a model with gating mechanisms can mitigate the errors and noise of individual modalities most effectively.

The LSTM model with gating outperforms other multimodal and single modality depression detection models in both binary and multi-class classification tasks. There is a significant performance boost by integrating textual and audio modalities with gating over not using it (F1 0.80 vs. 0.69; MAE 3.66 vs. 4.98). Adding visual features improves the performance despite the fact that word-alignment models cannot be easily used to combine frame-level visual information due to the fact the relatively slow frame rate from the visual information does not allow consistent overlap with the input word's duration (F1 0.81 vs. 0.69; MAE 3.61 vs. 3.66). The text features are highly informative for depression classification on their own, and without the appropriate fusion techniques the performance level can in fact decrease: integrating other modalities without gating control led to a slightly worse performance in our experiments (F1 scores 0.67 vs. 0.69; MAE 5.29 vs. 4.98).

In terms of our competitor baselines, while [18] and [14]'s multimodal classifiers performed better than all the unimodal

[1] Note this differs to [14] who found the interviewer's questions to contain highly predictive features.

Table 2: *Depression classification results using Unimodal features*

Model	F1	Prec.	MAE	RMSE
LSTM with Lexical Features	0.69	0.68	4.98	6.05
LSTM with Audio Features	0.66	0.71	5.21	6.44
LSTM with Visual Features	0.59	0.63	5.38	6.72

Table 3: *Depression classification results of systems with different fusion techniques*

Fusion Method	F1	Prec.	MAE	RMSE
Early Fusion	0.67	0.63	5.29	6.68
Late Fusion with Weighting	0.70	0.78	3.92	5.86
Model-Based Fusion	0.81	0.80	3.61	4.99

models, showing some useful fusion, we note that they both utilized utterance-level fusion and ignored the time-scale associations, meaning that these models may not function word-by-word incrementally. For integration into any live system, we suggest incremental processing is vital. Furthermore, our model outperforms models without utilizing the topic/context of questions and sequences of responses [12, 14] and the model with word-level audio features achieves better F1 and MAE performance in comparison to Alhanai et al. [18] that uses set of higher-order statistics as audio features for each individual's response. This indicates the potential advantages of an incremental word-level structure over employing global information across different time scales, without needing look-ahead for utterance-global or dialogue-global features. The model we proposed, utilizing sequence of utterances and trying to capture important temporal interactions, without conditioning on the topic of the query, performs better than [14]'s state-of-the-art baselines with context/topic modeling (F1 0.81 vs. 0.70 and MAE 3.61 vs. 3.84).

5.1. Fusion Analysis

Text is the most influential modality in detecting depression in a word-level structure in this dataset. From Table 2, we can see the performance of our LSTM models across modalities. Using only the text modality gives a better depression prediction than utilizing unimodal audio and visual modalities sequentially. Adding modalities to the LSTM with text without gating does not lead to improvement. Utilizing more modalities even results in worse performance in both MAE and F1 compared to unimodal LSTM with lexical features alone (Table 1). The audio and visual modalities can negatively impact the model's performance if word-level multimodal fusion is not controlled.

Our models, integrating multimodal features for each word, show improvement over Alhanai et al. [18] which attempted to find optimal input parameters for each modality, showing the potential advantages of a word-level time-dependent approach with effective fusion. When we employ gating, Table 1 indicates that more input modalities leads to better results in both F1 and MAE. We assume that the LSTM with gating succeeds in dealing with features in different contexts conveying different information at different rates and contributing different parts of the overall representation in the network. While the lexical content of the subjects' responses is clearly a strong indicator of depression in this dataset, the acoustic quality of each word is also indicative of depression, and visual information based on the bodily movement of the subject concurrent with their words also helps depression classification, albeit less markedly. While our simple technique of capturing information over word durations works well here, in future work we will explore more principled ways of capturing gesture/bodily movement data before its combination with lexical and acoustic data.

In terms of fusion techniques, the results in Table 3 show the model-based fusion method, designed to perform multimodal fusion within the network's architecture, obtains the highest performance. It benefits from observing temporal multimodal information and the ability to train both the multimodal representation and the fusion component simultaneously. The late fusion model performs better than the early fusion method (F1 0.70 vs. 0.67 and MAE 3.92 vs. 5.29) with the precision close to the model-based methods (Prec. 0.78 vs. 0.80). Late fusion approaches have the advantage of interpretability in terms of showing which modality is given the highest weight in the input, but they do not make use of the possible dependencies between modalities in real-time communication. Early fusion only needs one model for all modalities, making it the easiest and fastest method for training, however the network is not learning from the large heterogeneous input vector as effectively as the model-based version.

6. Conclusion

We have presented a model that learns the indicators of depression from audio and visual modalities as well as lexical information in transcript texts. We utilized word-level multimodal fusion with feed-forward highway layers as a gating mechanism. Our principal motivation is to capture inter-modal dynamics in a joint multimodal representation. Our model outperforms the state-of-the-art methods in both binary and numeric depression classification tasks.

In future work we intend to analyze the interactions between different modalities as the predictors of depression as they occur in real time. Monitoring the multimodal fusion after each word could help highlight informative moments that contribute more to the prediction of depression, which could in turn have several clinical applications for psychiatric practitioners in helping further understand symptoms of depression during interaction. Furthermore, we intend to undertake a more principled approach to the visual modality in terms of extracting bodily action sequences from motion capture data, which in turn interact with the verbal behaviour to give multimodal meaning.

7. Acknowledgements

We thank the reviewers for their helpful comments. Purver was partly supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The 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.

8. References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, p. e442, 2006.

[2] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depres-

sion," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[3] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[4] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[6] J. E. Perez and R. E. Riggio, "Nonverbal social skills and psychopathology," *Nonverbal Behavior in Clinical Settings*, pp. 17–44, 2003.

[7] J. T. M. Schelde, "Major depression: Behavioral markers of depression and recovery," *The Journal of Nervous and Mental Disease*, vol. 186, no. 3, pp. 133–140, 1998.

[8] P. Waxer, "Nonverbal cues for depression." *Journal of Abnormal Psychology*, vol. 83, no. 3, p. 319, 1974.

[9] C. Segrin, "Social skills deficits associated with depression," *Clinical Psychology Review*, vol. 20, no. 3, pp. 379–403, 2000.

[10] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 21–30.

[11] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 2013, pp. 160–169.

[12] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 69–76.

[13] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 53–59.

[14] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 11–18.

[15] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 35–42.

[16] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 61–68.

[17] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 43–50.

[18] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.

[19] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[22] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 163–171.

[23] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," *arXiv preprint arXiv:1807.06233*, 2018.

[24] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus."

[25] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[26] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[27] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[28] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[29] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt, "Screening for depression in medical settings with the patient health questionnaire (phq): a diagnostic meta-analysis," *Journal of General Internal Medicine*, vol. 22, no. 11, pp. 1596–1602, 2007.

[30] J. Pennington, R. Socher, and C. 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.

[31] 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.

[32] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[33] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[34] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Sweden, Tech. Rep. 4, 1985.

[35] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[36] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1491–1498.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Appendix F: A corpus study on questions, responses and misunderstanding signals in conversations with Alzheimer's patients

A Corpus Study on Questions, Responses and Misunderstanding Signals in Conversations with Alzheimer's Patients*

Shamila Nasreen, Matthew Purver, Julian Hough
Cognitive Science Group / Computational Linguistics Lab
School of Electronic Engineering and Computer Science
Queen Mary University of London Mile End Road,
London E1 4NS, UK
{shamila.nasreen,m.purver,j.hough}@qmul.ac.uk

Abstract

This paper describes an initial corpus study of question-answer pairs in the Carolina Conversations Collection corpus of conversational interviews with older people. Our aim is to compare the behaviour of patients with and without Alzheimer's Disease (AD) on the basis of types of question asked and their responses in dialogue. It has been suggested that questions present an interesting and useful phenomenon for exploring the quality of communication between patients and their interlocutors, and this study confirms this: questions are common, making up almost 14% of utterances from AD and Non-AD patients; and type distributions vary, interviewers asking many Yes-No questions (nearly 6%) from AD patients while more Wh-questions (5.4%) from Non-AD patients. We also find that processes of clarification and coordination (e.g. asking clarification questions, signalling non-understanding) are more common in dialogue with AD patients.

1 Introduction

Alzheimer's Disease (AD) is an irreversible, progressive deterioration of the brain that slowly destroys memory, language and thinking abilities, and eventually the ability to carry out the simplest tasks in patients' daily lives. AD is the most prevalent form of dementia, contributing to 60%-70% among all types of dementia (Tsoi et al., 2018). The most common symptoms of AD are memory lapses, difficulty in recalling recent events, struggling to follow a conversation, repeating the conversation, delayed responses, difficulty finding words for talk, and orientation problems (e.g. confusion and inability to track daily activities).

Diagnosis can be based on clinical interpretation of patients' history complemented by brain scanning (MRI); but this is time-consuming, stressful, costly and often cannot be offered to all patients complaining about functional memory. Instead, the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and Alzheimer's Association established criteria for AD diagnosis require the presence of cognitive impairment to be confirmed by neuropsychological testing for a clinical diagnosis of possible or probable AD (McKhann et al., 1984). Suitable neuropsychological tests include the Mini-Mental Status Examination (MMSE; Folstein et al., 1975, one of the most commonly used tests), Mini-Cog (Rosen et al., 1984), Addenbrooke's Cognitive ExaminationRevised (ACE-R; Noone, 2015), Hopkins Verbal Learning Test (HVLT; Brandt, 1991) and DemTect (Kalbe et al., 2004).

However, these tests require medical experts to interpret the results, and are performed in medical clinics which patients must visit for diagnosis. Currently, researchers are therefore investigating the impact of neurodegenerative impairment on patients' speech and language, with the hope of deriving tests which are easier to administer and automate via natural language processing techniques (see e.g. Fraser et al., 2016a).

In this paper, we focus on language in conversational interaction. We explore this as a diagnostically relevant resource to differentiate patients with and without Alzheimer's Disease (AD vs. Non-AD), using the Carolina Conversations Collection data in which patients interact with researchers and community persons on different but not prefixed topics like discussion about breakfast, lunch, special occasions (thanksgiving, Christ-

*This research was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). 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.

mas) etc. We particularly focused on the types of questions asked from both groups, how they are responded to, and whether there are any significant patterns that appear to differentiate the groups.

The remainder of this paper is organized as follows. In the next section, we describe earlier work on analyzing conversational profiles of AD and particularly on the types of questions they focused on. In Section 3 we give details about our new corpus study and annotation scheme. We then present and discuss the results in Section 4: in particular, how the distributions of different types of questions, and frequencies of signals of non-understanding, clarification questions and repeat questions, differ between AD patients and Non-AD. We close with a discussion of the overall result, and of possible further work.

2 Related Work

Recent years have seen an increasing amount of research in NLP for dementia diagnosis. Much of this work has looked at properties of an individual's language in isolation: acoustic and lexical features of speech (Jarrold et al., 2014), or syntactic complexity, fluency and information content (Fraser et al., 2016b,a). However, this is usually studied within particular language tasks, often within specific domains (e.g. the Cookie Theft picture description task of the DementiaBank Pitt Corpus[1]); however, conversational dialogue is the primary area of human natural language use, and studying the effects of AD on dialogue and interaction — and particularly more open-domain dialogue — might therefore provide more generally applicable insights.

Recent statistical modelling work shows that AD has characteristic effects on dialogue. Luz et al. (2018) extract features like speech rate, dialogue duration and turn taking measures, using the Carolina Conversations Collection corpus (Pope and Davis, 2011) of patient interview dialogues, and show that this can build a predictive statistical model for the presence of AD.

Work in the conversation analysis (CA) tradition has looked in more detail at what characteristics of dialogue with dementia might be important. Jones et al. (2016) present a CA study of dyadic communication between clinicians and patients during initial specialist clinic visits, while Elsey et al. (2015) highlighted the role of carer, looking at triadic interactions among a doctor, a patient and a companion. They establish differential conversational profiles which distinguish between non-progressive functional memory disorder (FMD) and progressive neuro-degenerative Disorder (ND), based on the interactional behavior of patients responding to neurologists' questions about their memory problems. Features include difficulties responding to compound questions, providing specific and elaborated examples and answering questions about personal information, time taken to respond and frequent "I don't know" responses.

Questions present an interesting testing ground when exploring the quality of communication between caregivers and persons with AD. Question-answer sequences have long been seen as a fundamental building block of conversation; Sacks et al. (1978) formalized these as a type of adjacency pair in which the first utterance represents the question and the second one is an answer. Hamilton (2005) explored the use of questions in conversation with a patient of AD over a period of four years, finding that Yes-No questions are responded to much more frequently than open-ended question i.e Wh-questions. Gottlieb-Tanaka et al. (2003) used a similar approach, examining Yes-No and open-ended questions in a conversation between family caregivers and their spouse with AD during different activities of daily life. They reported that caregivers used YesNo questions much more frequently than open-ended questions (66% vs. 34%, respectively) and there are fewer communication breakdowns with Yes-No Questions.

Varela Suárez (2018) worked specifically to observe dementia patients' ability to respond to different types of questions including close-ended questions, open-ended questions, and multiple choice questions. The objective of this study was to verify a) if the ability to answer questions persists until the final stages of dementia b), check if the number of preferred and relevant answers decreases progressively. The interviewers had a list of questions about patients memories, experiences, and daily routine, and were told to talk on the topics introduced by the patients, and only ask the questions from the list when patients are silent. The basic Question-Answer adjacency pair is preserved until the severe stage of the disease; however, the number of answered questions, preferred and relevant answers starts to decrease.

[1] http://talkbank.org/DementiaBank/

These studies show that the presence of AD affects the production of questions, their use and their responses, but all focus on specific types of question including Yes-No, Wh-questions, and Multiple choice questions. As far as we are aware, none of these studies have extended this approach to look into specific aspects of non-understanding or inability to respond: e.g. non-understanding signals, clarification requests and repetition of questions.

Dialogue Act Models

The ability to model and detect discourse structure is an important step toward working spontaneous dialogue and the first analysis step involves the identification of Dialogue Acts (DAs). DAs represent the meaning of utterances at the level of illocutionary force (Stolcke et al., 2000). Classifying utterances and assigning DAs is very useful in many applications including answering questions in conversational agents, summarizing meeting minutes, and assigning proper DAs in dialogue based games. DAs tagsets classify dialogue utterances based on the syntactic, semantic and pragmatic structure of the utterance.

The most widely used dataset and tagset in DA tagging is the Switchboard corpus, consisting of 1155 annotated conversations containing 205K utterances, 1.4 million words from 5 minute recorded telephonic conversations. The DA types and complete tagset can be seen in (Jurafsky et al., 1997). The corpus is annotated using a variant of the DAMSL tagset (Core and Allen, 1997) with approximately 60 basic tags/classes which combines to produce 220 distinct labels. Jurafsky et al. (1997) then combine these 220 labels into 43 major classes including *Statements, Backchannels, Questions, Agreements, Apology etc*.

3 Material and Methods

3.1 Research Questions

This study is a part of a larger project where we analyze what are the significant key indicators in the language and speech of AD patients that can be used as Bio-Markers in the early diagnosis process of Alzheimer's Disease. The focus of the initial and current study is on the interaction of AD patients and Non-AD patients with interviewers.

Our account suggests these interactions are based on what is being asked from the AD and Non-AD sufferers.We hypothesize that the distribution of questions being asked and the responses generated are not same for both the groups. We hypothesize that the use of different question types such as binary yes-no questions (in interrogative or declarative form), tag questions, and alternative ('or') questions will differ between groups; and the signals of non-understanding, back-channels in question form and clarification requests should be more common with AD patients.

In more detail, we are conducting this corpus study to answer the following research questions:

Q 1 *Is the distribution of question types asked by the patient and interviewer different when the patient is an AD sufferer?*
Our first interest is in the general statistics regarding what types of questions are asked of the AD and non-AD group. How often does each type occur, and what is the balance between the two groups? What types of questions are more frequently asked from Alzheimer's patients?

Q 2 *How often do signals of non-understanding, clarification requests and back-channel questions occur in dialogues with an AD sufferer compared to those without one?*
We hypothesize that due to the nature of AD, there will be more non-understanding signals and clarification questions in response to questions and statements.

Q 3 *Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?*
We hypothesize that there will be more repeated questions for the AD group from the interviewer, as AD patients find it difficult to follow a conversation.

3.2 Corpus

Our intention was to investigate the behavior of AD patients on the basis of questions and responses observed in a corpus of dialogue. For this purpose, we used the Carolina Conversation Collection (CCC), collected by the Medical University of South Carolina (MUSC)[2] (Pope and Davis, 2011). This dataset comprises of two cohorts: cohort one contains 125 unimpaired persons of 65

[2] https://carolinaconversations.musc.edu/

years and older with 12 chronic diseases with a total of 200 conversations. Cohort two includes 400 natural conversations of 125 persons having dementia including Alzheimer's of age 65 and above who spoke at least twice annually with linguistic students. The demographic and clinical variables include: age range, gender, occupation prior to retirement, diseases diagnosed, and level of education (in years) are available. As this dataset includes only older patients with diagnosed dementia, it can only allow us to observe patterns associated with AD at a relatively advanced stage, and not directly tell us whether these extend to early stage diagnosis. However, it has the advantage of containing relatively free conversational interaction, rather than the more formulaic tasks in e.g. DementiaBank. Work in progress is collecting a dataset of conversational language including early-stage and un-diagnosed cases; until then we believe this to be the most relevant corpus for our purposes.

The dataset consists of audio, video and transcripts that are time aligned. The identity of patients and interviewer is anonymized keeping in mind security and privacy concerns. Online access to the dataset was obtained after gaining ethical approval from Queen Mary University of London (hosting the project) and Medical University of South Carolina (MUSC, hosting the dataset), and complying with MUSC's requirements for data handling and storage.

For our corpus analysis here, we used dialogue data from 10 randomly sampled patients with AD (7 females, 3 males) and 10 patients with other diseases including diabetes, heart problems, arthritis, high cholesterol, cancer, leukemia and breathing problems but not AD (8 females, 2 males). These groups are selected to match age range, to compare the different patterns of interaction and to avoid statistical bias. This portion comprises of 2554 utterances for the AD group and 1439 utterances for the Non-AD group, with a total of 3993 utterances from 20 patients with 23 dialogue conversations.

The CCC transcripts are already segmented at the utterance (turn) level and the word level, and annotated for speaker identity (patient vs. interviewer); however, no DA information is available. We used only the utterance level layers; transcripts were available in ELAN format and we converted them to CSV format. We then manually annotate the transcripts at the utterance level with DA information.

3.3 Terminology

Throughout this paper, we use specific terms for particular question types and response types, and use these in our annotation procedure. Following Switchboard's SWBD-DAMSL terminology (Jurafsky et al., 1997), we use **qy** for **Yes-No** questions, and **qy^d** for **Declarative Yes-No** questions. Declarative questions (^d) are utterances which function pragmatically as questions but which do not have "question form" in their syntax. We use **qw** for **Wh-questions** which includes words like *what, how, when, etc.* and **qw^d** for **Declarative Wh-**questions. Yes-No or Wh-questions are questions which do not have only pragmatic force but have a syntactic and prosodic marking of questions or interrogative in nature. We used **^g** for **Tag questions**, which are simply confirming questions that have auxiliary inversion at the end of statement e.g. (*But they're pretty, aren't they?*). For **Or questions** which are simply choice question and aids in answering the question by giving choices to the patients are represented by **qr** e.g (- *did he um, keep him or did he throw him back?*).

We used term **Clarification question** for questions that are asked in response to a partial understanding of a question/statement and are specific in nature. These clarification questions are represented by **qc**. **Signal non-understanding** is generated by a person in response to a question that they have not understood and are represented by **br**. **Back-channel Question (bh)** is a continuer which takes the form of question and have question intonation in it. Back-channels are more generic than clarification questions and often occur in many types (*e.g really? Yeah? do you? is that right? etc.*).
When the response to a Yes-No question is just a yes including variations (e.g. *yeah, yes, huh, yes, Yes I do etc.*), it will be represented by **ny** and when there is a yes plus some explanation, it will be represented by **ny^e**.

(1) A: Do you have children?
B: Yeah, but they're big children now. Grown.

[CCC Mason_Davis_001 28-29]

na is an affirmative answer that gives an explanation without the yes or its variation. **nn** is used for

No-answers and **nn^e** is used for an explanation with No answer (see Appendix A for Examples).

3.4 Annotation Scheme

The original SWBD-DAMSL tagset for the Switchboard Corpus contains 43 DA tags (Jurafsky et al., 1997). Our initial manual includes DA tags from SWBD-DAMSL and our own specific new DA tags with a total of 35 tags. For different types of questions and their possible responses, 14 DA tags are taken from SWBD-DAMSL and 2 new tags are introduced. These new tags are for clarification questions **(*qc*)** and for answers to Wh-Questions **(*sd-qw*)**, and were required to distinguish key response types.[3]

The ability to tag specific clarification questions is important for our study, as questions asked by the interviewer can be followed by a clarification which indicates partial understanding while requesting specific clarifying information (SWBD-DAMSL only provides the ***br*** tag for complete non-understanding). The distinction between answers to Wh-Questions and other, unrelated statements is also important (in order to capture whether the response is relevant: a relevant answer should be different from simple general statement), but SWBD-DAMSL provides only a single ***sd*** tag for statements. Different types of question and their tags are given with examples in Table 1; a list of response types is given in Table 2.

Another new addition is the tagging of *repetition* of questions, with or without reformulation. We marked repeat questions as simple repeats or reformulations, and tagged with the index of the dialogue act (utterance number) they were repeating or reformulating.

Similarly, clarification questions can signal non-understanding with two main distinct CR forms, and this distinction is tagged: pure repeats and reformulated repeated questions that are slightly changed syntactically but the context remains the same – see Table 3 with utterance 144.

3.5 Inter-Annotator Agreement

To check inter-annotator agreement, three annotators annotated one conversation of an AD patient and Non-AD interviewer of 192 utterances. All annotators had a good knowledge of linguistics and were familiar with both the SWBD-DAMSL tagset and the additions as specified above and in the manual. First, all three annotators annotated the dialogue independently by assigning DA tags to all utterances with the 17 tags of interest for this paper as shown in Table 4 ('other' means the annotator judged another SWBD-DAMSL act tag could be appropriate apart from the 16 tags in focus). We use a multi-rater version of Cohen's κ (Cohen, 1960) as described by (Siegel and Castellan, 1988) to establish the agreement of annotators for all tags and also 1-vs-the-rest as shown in Table 4 below.[4]

As can be seen, an overall agreement was good (κ=0.844) for all tags and the majority of tags which were tagged by any annotator in the dialogue have $\kappa > 0.67$, with only 'no' getting beneath $\kappa < 0.5$. We judged this test to be indicative of a reliable annotation scheme for our purposes.

4 Results and Discussion

From the CCC transcripts, we selected 23 conversations, which when annotated yield 3993 utterances. All utterances were tagged with one of the 16 dialogue act tags relating to all question categories and their possible answers as described above, plus an 'other' tag. In addition to the dialogue act tag, utterances deemed to be responses (tags in Table 2) were tagged with the index of the utterance being responded to. Repeat questions were also marked as *simple repeats* or *reformulations*, and tagged with the index of the dialogue act they were repeating or reformulating.

Is the distribution of question types asked by the patient and interviewer different when the patient is an AD sufferer?

To investigate the distribution of dialogue acts, we calculated the relative frequency of each question and response type separately for AD and Non-AD group, and for the patient and interviewer within those groups. A comprehensive analysis of particular types and their distribution between AD and Non-AD patient with their interviewer is shown in Table 5. More yes-no questions (qy) are asked by the interviewer from AD Patients than Non-AD patients (6% vs 3.7%) and fewer wh-questions

[3] Some other DA tagging schemes provide categories for these and more; however, we chose to begin with SWBD-DAMSL given its prevalence in DA tagging work, and extend it only as necessary. In future work we plan to examine multi-dimensional schemes (e.g. Core and Allen, 1997; Bunt et al., 2010) to see if they provide benefits in this setting.

[4] The annotation results and scripts are available from https://github.com/julianhough/inter_annotator_agreement.

Type	Tag	Example
Yes-No Question	qy	Did you go anywhere today?
Wh-Question	qw	When do you have any time to do your homework?
Declarative Yes-No Question	qy^d	You have two kids?
Declarative Wh-Question	qw^d	Doing what?
Or Question	qr	Did he um, keep him or did he throw him back?
Tag Question	^g	But they're pretty aren't they?
Clarification Question	qc	Next Tuesday?
Signal Non-understanding	br	Pardon?
Backchannel in question form	bh	Really?

Table 1: Question Types for CCC

Type	Tag	Example
Yes answer	ny	Yeah.
Yes- plus expansion	ny^e	Yeah, but they're
Affirmative non-yes answer	na	Oh I think so. [laughs]?
No answer	nn	No
Negative non-no answers	nn^e	No, I belonged to the Methodist church.
Other answer	no	I, I don't know.
Declarative statement wh-answer	sd-qw	Popcorn shrimp and it was leftover from yesterday.

Table 2: Answer Types for CCC

Tag	Speaker:Utterance	Text	Repeat Question?
qw	A:15	-Where's she been?	
br	B:16	-Pardon?	
qw	A:17	-Where is she been?	15
qy	A:142	-Well, are you, are you restricted from certain foods?	
br	B:143	-What?	
qy	A:144	-Like, do they, do they make you eat certain foods because your medication?	142-reformulation

Table 3: Examples of Repeated questions

(qw) are asked in the AD group compared to the non-AD group (4% vs 5.4%). Choice questions (qr) are also asked more from AD patients compared to non-AD patients (2% vs 0.3%). These results suggest there is a systematic difference in question distributions; one plausible explanation for this is that AD patients find it easier to answer a simple Yes-No question or a choice question compared to a wh-question. It is also obvious from the results that AD patients are also asking more questions than Non-AD patient during their conversation with the interviewer (*qy*: 1% vs 0.3%), (*qw*: 1% vs 0.3%), (*^g*: 0.2% vs 0.1%), (*br*: 3% vs 0.4%), and (*qc*: 2% vs 0.1%).

We also compared the distribution of these tags with the Switchboard SWDA corpus, as shown in Table 6. As the CCC is a set of clinical interviews, the percentage of tags which are questions is higher in this corpus compared to Switchboard. Although simple yes-no questions have almost identical frequencies in both corpora, declarative yes-no, wh-questions, declarative wh-questions, tag questions, and signals of non-understanding are higher in the CCC than Switchboard. Our new clarification question (qc) tag accounts for 1% for both AD group and Non-AD group tags but is not annotated in SWDA.

Tag	# times annotated	κ
qy	26	0.758
qw	30	0.895
qy^d	12	0.660
qw^d	3	1.000
^g	2	0.498
br	22	0.953
bh	0	0
qc	15	0.795
qr	0	0
ny	12	1.000
ny^e	11	0.907
na	8	0.873
nn	1	0
nn^e	6	0.663
no	4	0.497
sd-qw	26	0.637
other	398	0.902
all tags	576	0.844

Table 4: Multi-rater Cohen's κ statistics for one-vs-rest and overall agreement score for one dialogue.

DA tag	AD		Non-AD	
	Pat	Int	Pat	Int
qy	1%	6%	0.3%	3.7%
qy^d	1%	6%	0.1%	5%
qw	1%	4%	0.1%	5.4%
qw^d	0.4%	1%	0.5%	0
^g	0.2%	2%	0.1%	0.7%
qr	0.1%	2%	0	0.3%
br	3%	0.1%	0.4%	0
bh	1%	1%	1%	1%
qc	2%	1%	0.1	1%
simple-Repeat	0	1%	0	0
reformulation	0	2%	0	0

Table 5: Distribution of DA question tags among the AD group and Non-AD group

DA Tag	CCC-AD	CCC-Non-AD	SWDA
qy	3%	2%	2%
qy^d	4%	2%	1%
qw	3%	3%	1%
qw^d	1%	0.3%	<.1%
^g	1%	0.5%	<.1%
br	1%	0.2%	0.1%
bh	1%	1%	1%
qc	1%	1%	-
qr	1%	0.2%	0.1%
ny	3%	1%	1%
ny^e	2%	2%	0.4%
na	3%	3%	1%
nn	0.4%	0.4%	1%
nn^e	1%	1%	0.1%
no	0.4%	0.3%	1%
sd-qw	4%	6%	-

Table 6: Comparison of relative frequency of DA tags in the AD group, Non-AD group of the CCC and SWDA corpora

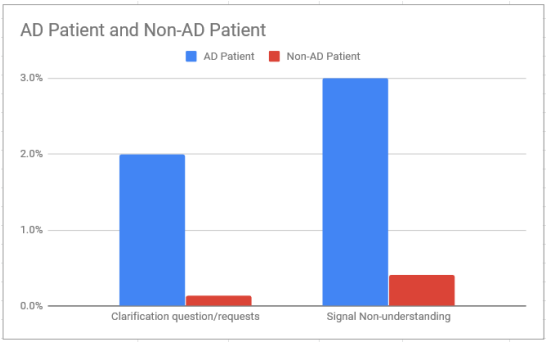

Figure 1: Clarification questions and Signal Non-understanding

How often do signals of non-understanding, clarification requests and back-channel questions occur in dialogues with an AD sufferer compared to those without one?

An examination of signals of non-understanding, clarification requests and back-channel requests reveals that the ability to follow and understand questions decrease for AD patients so they produce more signals of non-understanding (e.g *sorry Maam?, Pardon?, huh?, eh?*), when questions are posed to them. On the other hand, signals of non-understanding from Non-AD patients are much less frequent as shown in Figure 1. The overall frequency of clarification questions (qc) between the two conversation groups was not systematically different as shown in Table 6 when utterances from both patient and interviewer are combined, but dealing with them separately, AD patients produce more clarification requests than non-AD patients (2% vs 0.1%) – see Table 5 and Fig. 1.

We further examine how often signals of non-understanding and clarification requests are issued in response to questions rather than statements/answers. Examination of the data shows that clarification requests are more often gener-

	AD Group	Non-AD Group
Question followed by Signal of Non-understanding	24 (35)	2 (3)
Statements followed by Signal of Non-understanding	11 (35)	1 (3)
Question followed by Clarification Question	8 (34)	1 (11)
Statement followed by Clarification Question	26 (34)	10 (11)

Table 7: Occurrences of signal non-understanding and clarification question followed by question/statements

ated in response to statements, and less often after questions are raised; but signal non-understanding happen more often after questions. Out of total 35 signal non-understanding, 24 are generated in response to a question of AD Group as shown in Table 7. However, only 8 clarification questions are asked in response to questions, with 26 asked in response to declarative statements – (see Appendix A for more examples and context).

Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?

Many questions are followed by clarification questions or signal non-understanding, so there will be more repetition of a similar type of question in case of the AD patients. Repeated questions are asked in two variations; either repeated simply or reformulated so that the patient can understand the question properly as in (4). In the AD group 4.7% questions are simple-repeat questions and 6.7% are reformulated as shown in Table 8 while for the non-AD group only 2.4% are reformulated questions and there were no repeated questions.

(4) A: Your dad worked for who was it? Swisten
A: and that's why you went up to Baltimore?.
B: Huh?
A: Your dad went to –worked at – worked for Swisten?
B: My Father?
A: Yeah. Is that why you guys went to Baltimore?
[CCC Tappan_Patte_001 37-43]

Repeat Type	AD Group	Non-AD Group
Total Question	313	127
Simple-Repeat Question	15 (4.7%)	0
Reformulated Question	21 (6.7%)	3 (2.4%)

Table 8: Repetition and reformulation of questions for AD group and Non-AD group

5 Conclusion and Future work

Our study provides the first statistical analysis of different types of question asked in conversations with AD patients in the Carolina Conversation Collection (CCC) Corpus. We found that yes-no questions were asked more frequently in the AD sufferer conversations than the Non-AD conversations (6% vs 3.7% of all dialogue acts) and less Wh-questions were asked in AD sufferer conversations compared to Non-AD ones (4% vs 5.4%). While our newly introduced tags were not frequent, they are significant in AD sufferer conversations, with 2% of all dialogue acts by AD sufferers being clarification questions and 3% being signals of non-understanding.

In future work, we plan to work on the CCC corpus conversations of both AD and Non-AD conversations to build an automatic dialogue act tagger for the tagset we used in this study. We will also explore more complex questions including compound questions and questions that relate to semantic memory and episodic memory. We also plan to look into disfluency and repairs in this data collection which could further aid interpretation and automatic diagnosis.

References

Jason Brandt. 1991. The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5(2):125–142.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation*.

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

Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56. Boston, MA.

Christopher Elsey, Paul Drew, Danielle Jones, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, Annalena Venneri, and Markus Reuber. 2015. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077.

M F Folstein, S E Folstein, and P R McHugh. 1975. Mini-mental status. a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016a. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Kathleen C. Fraser, Frank Rudzicz, and Graeme Hirst. 2016b. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In *Proc. CLPsych*, pages 1–11, San Diego, CA, USA. Association for Computational Linguistics.

Dalia Gottlieb-Tanaka, Jeff Small, and Annalee Yassi. 2003. A programme of creative expression activities for seniors with dementia. *Dementia*, 2(1):127–133.

Heidi Ehernberger Hamilton. 2005. *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proc. CLPsych*, pages 27–37, Baltimore, Maryland, USA. Association for Computational Linguistics.

Danielle Jones, Paul Drew, Christopher Elsey, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, and Markus Reuber. 2016. Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. *Aging & Mental Health*, 20(5):500–509.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual.

Elke Kalbe, Josef Kessler, Pasquale Calabrese, R Smith, AP Passmore, Met al Brand, and R Bullock. 2004. DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *International journal of geriatric psychiatry*, 19(2):136–143.

Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. In *Proceedings of the LREC 2018 Workshop Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)*.

Guy McKhann, David Drachman, and Marshall Folstein. 1984. Clinical diagnosis of Alzheimer's disease. *Neurology*, 34(7):939—-944. Views & Reviews.

Peter Noone. 2015. Addenbrooke's Cognitive Examination-III. *Occupational Medicine*, 65:418–420.

Charlene Pope and Boyd H Davis. 2011. Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

W G Rosen, R C Mohs, and K L Davis. 1984. A new rating scale for Alzheimer's disease. *American Journal of Psychiatry*, 141(11):1356–1364.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier.

Sidney Siegel and NJ Castellan. 1988. Measures of association and their tests of significance. *Nonparametric Statistics for the Behavioral Sciences*, pages 224–312.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Kelvin K F Tsoi, Lingling Zhang, Nicholas B Chan, Felix C H Chan, Hoyee W Hirai, and Helen M L Meng. 2018. Social Media as a Tool to Look for People with Dementia Who Become Lost : Factors

That Matter. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 9:3355–3364.

Ana Varela Suárez. 2018. The question-answer adjacency pair in dementia discourse. *International Journal of Applied Linguistics*, 28(1):86–101.

A Examples from Carolinas Conversation Collection

Yes-No Question followed by no plus expansion answer:

Tag	Text
qy	A:were you Primitive Baptist?
nn^e	B: — no, I belonged to the Methodist church.

[CCC Mason_Davis_001 92-93]

Yes-No Question followed by other answer:

Tag	Text
qy	A: are you going to go with them to see the Christmas Lights?
no	B: Oh, I, I dont know.

[CCC Wakefield_Brock_001 51-52]

Two Wh-Questions followed by declarative statements wh-answer:

Tag	Text
qw	A: - what does he preach about?
sd-qw	B: – hell hot and heaven beautiful.
qw	C:what types of food do you like the best?
sd-qw	D – vegetables, meat,
+	- and desserts.

[CCC Mason_Davis_001 31-32]
[CCC Wakeman_Rhyne_001 6-7]

Wh-question followed by a clarification question(*qc*) and a wh-question followed by a statement and then a clarification(*qc*):

Tag	Text
qw	A: where is Jerusalem Primitive Baptist Church?
qy	- is that near Fountain Hill?
br	B: - m'am?
qw	A: where is that church?
qc	B: Fountain Hill?
qw	A: what do you do?
sd-qw	B - I'm a teacher.
qc	A: Preacher?-

[CCC Mason_Davis_001 83-86,64-66]

Declarative wh-question followed by signal non-understanding(*br*) and then by reformulated-repeat wh-question:

Tag	Text
qw^d	A: You were married for–
br	B: Huh?
qw	A: How long– have you been married? (*reformulated-repeat*)

[CCC Tappan_Patte_001 7-9]

Declarative statement followed by back-channel question(*bh*) and then by yes answer:

Tag	Text
sd	A: huh, it used to be something special. it used to be my Mother's birthday.
bh	B: Really ?
ny	A: Yeah

[Wheaden_Lee_001 52-54]

Appendix G: Interaction Patterns in Conversations with Alzheimer's Patients

Interaction Patterns in Conversations with Alzheimer's Patients*

Shamila Nasreen[1], Matthew Purver[1,2], and Julian Hough[1]

[1] Cognitive Science Group / Computational Linguistics Lab
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
[2] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Introduction This poster describes an ongoing study into interaction patterns in spontaneous spoken conversations of patients with Alzheimer's Disease (AD). Our aim is to compare the behaviour of patients with and without AD on the basis of a range of dialogue phenomena. Building on our previous work [4], we analyse the distributions of these phenomena in the Carolina Conversations Collection (CCC) corpus, comparing patients with and without AD, and find several differences which can help distinguish the two.

Method and Results In previous work [4] we showed that the distribution of question types asked by both parties in a dialogue differs between AD and Non-AD patients, with differences in the frequency of closed and open questions asked, and in the clarification and non-understanding behaviour which can follow (agreeing with similar findings using different methods and a different clinical setting e.g. [1,2]). In this paper, we investigate further phenomena, looking at the types of *responses* to questions asked, and at *delayed* responses via the presence of pauses, both within a single speaker turn and at speaker transition points.

We find that response types differ significantly, with AD patients more likely to answer yes-no questions positively, and Non-AD patients more likely to answer negatively, optionally expanding their answers. Pauses at speaker transition points, both from patient to interviewer and from interviewer to patient are found to differ significantly between the two groups, both in terms of number and duration (see also [3]). We hypothesize that these pauses reflect difficulty answering questions: this causes delays in transition from interviewer to patient after a question is asked, and delays in transition from patient to interviewer when a question is answered insufficiently well. Similarly, the difference in response types may be evidence of strategies on the part of AD patients to avoid complex answers or open discussion - see also [1].

This study confirms that these interaction patterns may serve as an index of internal cognitive processes that help in differentiating AD patients and Non-AD patients and may be used as an integral part of language assessment in clinical settings.

* This research was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). 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.

Tag	κ
qy	0.76
qw	0.89
qy^d	0.66
qw^d	1.00
^g	0.49
br	0.95
qc	0.79
qr	0
ny	1.00
ny^e	0.91
na	0.87
nn^e	0.66
no	0.50
sd-qw	0.64
other	0.90
all tags	0.84

Table 1. Multi-rater Cohen's κ score.

References

1. Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., Reuber, M.: Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. Patient Education and Counseling **98**(9), 1071–1077 (2015)
2. Jones, D., Drew, P., Elsey, C., Blackburn, D., Wakefield, S., Harkness, K., Reuber, M.: Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. Aging & Mental Health **20**(5), 500–509 (2016)
3. Luz, S., de la Fuente, S., Albert, P.: A method for analysis of patient speech in dialogue for dementia detection. In: Kokkinakis, D. (ed.) Proceedings of the LREC 2018 Workshop Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2) (2018)
4. Nasreen, S., Purver, M., Hough, J.: 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. SEMDIAL, London, United Kingdom (Sep 2019), http://semdial.org/anthology/Z19-Nasreen_semdial_0013.pdf