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

D3.1: Datasets, benchmarks and evaluation metrics for cross-lingual user generated content filtering and analysis (T3.4)

Executive summary

This report details the datasets and other resources identified to support the tasks dedicated to cross-lingual **user-generated content (UGC)** analysis within WP3 of the EMBEDDIA project. It describes the new datasets produced within the project, together with other existing relevant available resources, which are either public or can be shared within the consortium. The report concludes with a description of benchmarks and evaluation metrics expected to be used for WP3 tasks.

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:	**D3.1**
Deliverable full title:	**Datasets, benchmarks and evaluation metrics for cross-lingual user generated content filtering and analysis**
Deliverable short title:	**Datasets and evaluation for UGC**
Document identifier:	**EMBEDDIA-D31-DatasetsAndEvaluationForUGC-T34-submitted**
Lead partner short name:	**QMUL**
Report version:	**submitted**
Report submission date:	**30/09/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Matthew Purver (QMUL), Ravi Shekhar (QMUL)**
Co-author(s):	**(N/A)**
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
03/07/2019	v0.1	Matthew Purver (QMUL)	Initial version.
17/09/2019	v0.2	Matthew Purver, Ravi Shekhar (QMUL)	Initial draft for review.
22/09/2019	v0.3	Jose Moreno (ULR)	Internal review.
23/09/2019	v0.4	Shane Sheenan (UE)	Internal review.
24/09/2019	v1.0	Matthew Purver (QMUL)	Updates after internal review.
24/09/2019	v1.1	Nada Lavrač (JSI)	Report quality checked and finalised.
25/09/2019	final	Matthew Purver (QMUL)	Changes integrated with new section 4.2 added, report finalised.
30/09/2019	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

DNN	Deep Neural Network
NLP	Natural Language Processing
NER	Named Entity Recognition
SUD	Socially Unacceptable Discourse
UGC	User-Generated Content
WP	Work Package

1 Introduction

This report is part of the activities within WP3 of the EMBEDDIA project. The overall objective of WP3 is to use EMBEDDIA's cross-lingual technologies to develop tools to understand the reactions of multilingual news audiences, thus helping news media companies to better serve their audience, and acting as a basis to assure fairness and integrity of participants in public internet spaces. WP3 will produce a range of tools for automatic analysis of **user-generated content (UGC)**, specifically, the reader comments below news articles published online. We expect these to include tools to analyse the sentiment and opinion expressed in comments, to detect and filter inappropriate comments such as those expressing hate speech and abuse, and to summarise the results in easily comprehensible reports. The specific objectives of WP3 are as follows:

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

These objectives are pursued within tasks T3.1–T3.3 respectively, while the aim of task T3.4 is to collect and prepare public and private resources, namely the datasets required to develop and evaluate the monolingual and cross-lingual UGC classifiers, the metrics required to quantify those evaluations, and the benchmarks required to compare to the state of the art. This report describes the results of the work performed in T3.4 in the first nine months of project duration: identifying and gathering datasets for the UGC tasks of interest, and for the EMBEDDIA project languages (English, Slovene, Croatian, Estonian, Lithuanian, Latvian, Russian, Finnish, and Swedish) where possible.

The repository of collected training and evaluation data is stored in a dedicated private cloud available to all project partners. Many datasets have been collected from sources which are already publicly available or available for research purposes, and our purpose here is merely to collect them for project use. Some new datasets have been produced as part of the EMBEDDIA project, with annotation and augmentation ongoing, and these form the main testbed for our technologies as applied to our media partners' domains and needs. Our purpose is not only to use them to develop and evaluate the EMBEDDIA technology, but to make them available to enable wider future research, and we give details below where this is the case.

This report is split into further four sections. Section 2 describes the main analysis tasks which we expect to address in our UGC content analysis technologies, and which our collected datasets therefore characterise. Section 3 describes the available datasets and benchmarks, with Section 3.1 describing the existing datasets available from project-external public sources, and Section 3.2 the new datasets being generated as part of the EMBEDDIA project itself. Section 4 describes the evaluation metrics used for each task, and the benchmarks we currently take for comparison to state of the art performance. Section 5 presents some overall conclusions and outlines plans for further work, emphasising the incremental nature of dataset collection and benchmarking.

2 Analysis tasks

The high-level task structure of WP3 specifies the technologies that will be developed for UGC analysis in general terms: analysis tools in T3.1, filtering tools in T3.2 and reporting tools in T3.3. However, each of these will be founded on (and evaluated over) a range of more specific analysis and classification tasks. Many such tasks are relevant for UGC in the news media domain, and resources will be required to support each. As identified in the EMBEDDIA user needs workshop and detailed in Deliverable D6.3 (*Report on user needs and challenges for news media industry*), for the EMBEDDIA news media partners the main requirements centre around (a) the detection of comments that should be blocked or referred to moderators before appearing publicly, (b) the detection of undesirable behaviour such as trolling, and (c)

the analysis and use of useful comments and positive behaviour in understanding reactions and creating new content. Here, we briefly discuss how these general task areas relate to specific subtasks that we expect to approach in the first parts of the project, especially to tasks already defined and studied in the state of the art. We do not intend this to be an exhaustive list at this stage; other tasks may emerge as being important as research progresses.

2.1 Abuse, hate speech and offensive language

The detection of abusive or offensive language has seen a great deal of attention in recent years, with many public datasets and shared tasks released and many classification systems developed and tested. The exact definition of the categories annotated in these tasks varies (see Schmidt and Wiegand, 2017, for a survey), but may include:

- Threats: hostile speech intended to threaten the addressee with violence or other negative effects;
- Abuse: personal insults directed at others, including 'flaming' or cyberbullying;
- Hate speech: personal attacks on the basis of religion, race, sex, sexuality etc.;
- Offensive content: the use of language which is in itself considered rude, vulgar or profane (including pornographic), even if not targeted at someone in particular.

These terms are often used interchangeably, with some (particularly *hate speech*) often used to cover multiple categories. Exact definitions of the individual categories also vary with task and dataset, so we do not attempt an exhaustive exposition here. As an illustrative example, Waseem and Hovy (2016) use an inclusive definition in order to gather a range of phenomena, defining one joint *hate speech* category on Twitter as a message that:

1. *uses a sexist or racial slur;*
2. *attacks a minority;*
3. *seeks to silence a minority;*
4. *criticizes a minority (without a well founded argument);*
5. *promotes, but does not directly use, hatespeech or violent crime;*
6. *criticizes a minority and uses a straw man argument;*
7. *blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims;*
8. *shows support of problematic hash tags. E.g."#BanIslam", "#whoriental", "#whitegenocide";*
9. *negatively stereotypes a minority;*
10. *defends xenophobia or sexism;*
11. *contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.*

On the other hand, Ljubešić et al. (2019) use a more restrictive set of definitions via a decision tree to separate out different kinds of *socially unacceptable discourse (SUD)* on Facebook into different categories:

Is this SUD aimed at someone's background?
YES: Are there elements of violence?
YES: background, violence
NO: background, offensive speech
NO: Is this SUD aimed towards individuals or other groups?
YES: Are there elements of violence?
YES: other, threat
NO: other, offensive speech
NO: Is the speech unacceptable?
YES: inappropriate speech
NO: acceptable speech

In all these variants, the task is usually defined as a classification task — detecting whether a given text should be classified as hate speech (or abuse, offensive language etc.) or not — although this may be set up as a binary or a multi-class classification problem depending on the definitions used. In our news media setting, the ability to detect UGC which is likely to offend readers, which might be intended to abuse or threaten others, and/or which might be illegal, is clearly a vital task for news publishers. We expect this to be one of the core components of Task T3.2 (comment filtering).

2.2 Spam detection

Another important task for UGC filtering and analysis is the detection of *spam*: comments which are off-topic, intended not to contribute to an ongoing conversation or relate to a given topic but rather to advertise, and/or to entice readers into clicking on a link either to generate revenue or for more nefarious purposes (e.g. 'phishing', attempting to gain access to personal information). The task is a variant of the familiar spam detection problem for email (see Caruana and Li, 2012, for a survey), but UGC and online comments have their own distinctive characteristics – see for example (Kantchelian et al., 2012) for application to comments in the blog domain, (Aiyar and Shetty, 2018) in the Youtube domain, and (Wu et al., 2018) for a survey of work in the Twitter domain. Again, this task is usually defined as a binary classification task; and is highly relevant for news media companies in order to prevent comments sections being taken over by irrelevant, offputting or dangerous content. We expect this to be another important component of Task T3.2 (comment filtering).

2.3 Trolling and incitement

A recent problem in many online contexts is the presence of *trolls* and *bots*: users who may be automated or semi-automated rather than human, and which behave in a disruptive and/or deceptive manner in order to influence discussion, spread propaganda and manipulate opinion or to incite extreme views and disrupt discussion (see e.g. Kim et al., 2019). The effects of such agents in social media and news article comments can be strong, with evidence that they have affected public opinion and outcomes of elections (Badawy et al., 2018). There is a connection with the *fake news* phenomenon, with many trolling accounts being used to spread false rumours and link to fake news. We expect this to be another core component of Task T3.2 (comment filtering).

In this case, although this can be approached in a similar classification manner to the tasks above, labelling texts as coming from trolls, the problem is more often seen as one of classifying user accounts rather than their individual text outputs. Methods used therefore often depend as much on the social network properties of user accounts as on the language they generate.

2.4 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 and Lee, 2008). Again, 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 one of the core foci of T3.1 (context and opinion analysis), and will then provide components for other tasks: detecting positive and negative comments will be a fundamental component of Task T3.3 (report generation), and detecting negative emotions will be a component of 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 and Battersby, 2012) for multi-class emotion detection in the Twitter domain, (Celli et al., 2016a) for opinion mining in online web comments (including news UGC).

2.5 Topic analysis

Analysis of texts to detect, classify or track the topical content they contain is another standard NLP text processing task (see e.g. Blei, 2012); in conversational language this is also linked to the task of segmenting discourse into different topical segments (Purver, 2011), with the analogue to this in text comments such as UGC being to group comments together into topical clusters (Aker et al., 2016a). While this task can be seen as a classification task in some cases (e.g. when a set of topics is known *a priori*, and documents must be categorised accordingly), in many cases it is approached as an unsupervised task in which a suitable set of topics must be inferred from a dataset without prior knowledge. In the latter case, evaluating performance can be a difficult and subjective task, and learning direct from other datasets is often unhelpful, so less resource-intensive methods are often used. We discuss this further below.

In the case of UGC for news media, both versions of the task may be appropriate: the ability to classify comments as being on-topic (e.g. relative to the article content) or off-topic will be important in Task T3.3 (comment filtering), and the ability to detect and track old and new topics as they emerge in comment

threads will be important in Task T3.4 (summarisation and report generation) – see e.g. (Aker et al., 2016a), (Aker et al., 2016b) for some discussion and examples.

2.6 Conversation structure 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 the tasks described above will depend on the ability to automatically detect this conversation structure and suitably model the ongoing context: for example, a comment C2 in which the author agrees with a previous comment C1 may be an example of hate speech if C1 is an example of hate speech; it may express a positive opinion or a negative opinion depending on the opinion expressed in C1; and it may contribute to different topics depending on the content of C1. Understanding agreement and disagreement relations has therefore proved to be important in previous work on summarisation of news comments (Barker and Gaizauskas, 2016), and on understanding opinions in online comments (Celli et al., 2016a). We therefore expect this task to be an important focus of development in Task T3.1, and form a key component of many tools developed in Tasks T3.2 and T3.3.

Characterisation of this task varies, with most approaches examining sub-tasks such as agreement detection or antecedent detection and seeing them as standard binary classification tasks (see e.g. Celli et al., 2016b); tree- or graph-based variants can also be used, requiring different approaches to annotation and evaluation.

2.7 Summarisation and reporting

This task will develop and implement methods for generating human-readable reports, in multiple languages, from the outputs of the analysis and filtering tools developed in Tasks T3.1 and T3.2, using the components which are outputs of the tasks described above. Approaches to summarisation are generally characterised as either *extractive* — classification-based approaches which choose particularly significant or summary-worthy passages of the original text to output — or *abstractive*, generating reports from some other derived representation rather than the original text itself. Both have shown success in news and comment summarisation (Riccardi et al., 2016), and methods for both will be developed in WP5; here, we will apply WP5's outputs to the particular setting of UGC, following e.g. (Barker and Gaizauskas, 2016). Extractive approaches usually require annotated dataset to learn a classification model from; abstractive approaches may not, often using hand-built templates or rules to generate structured output.

3 Datasets and benchmarks

This section gives details of the datasets identified from existing sources and being created as part of the EMBEDDIA project. Section 3.1 details the main existing datasets that have been identified from external sources and are (or can be) available to the project, that are relevant to the tasks outlined above and can be used for tool development and testing; these are primarily in well-resourced languages,

mainly English. Section 3.2 then details the datasets collected by the project partners, in the project languages, which will be used for final development and testing of our cross-lingual solutions.

3.1 Existing datasets

This section details the main existing datasets identified as relevant for supporting tool development for WP3 tasks.

3.1.1 Abuse, hate speech and offensive language

Many datasets are available for this broad category of tasks, with a number of public shared tasks having been run over the last few years. A helpful catalogue of relevant datasets is already available online at `http://hatespeechdata.com/`. The exact categories annotated vary; we give an indication of each in Table 1 below. In Table 1, we also include a hyperlink to a URL where the dataset can be downloaded if such a URL is available.

Most datasets are based on social media (SM) (mainly Twitter) messages, but some datasets specifically providing news UGC comments (NC) with annotations are available, mostly in English (e.g. YNACC from Yahoo News), but with one in Greek. A few are based on forum comments (FC) or on Wikipidia (Wiki) article comments.

In the less-resourced languages of the EMBEDDIA project, few datasets exist; we are aware of and have access to the FRENK dataset of Facebook post data for Slovene (Ljubešić et al., 2019) (this dataset is not publicly available, but project access has been obtained), but none for the other languages. Providing this is one of the primary outputs of our new datasets (Section 3.2 below).

We also note the existence of Hatebase:[1] a highly multilingual collection of crowdsourced social media posts; however, as its annotation is based only on submission by the public, and it contains no comparable non-abuse language, we consider it as a slightly different category to the main datasets detailed here. It may be useful to the project for discovery of key vocabulary terms.

Table 1: Primary existing datasets: abuse, hate speech and offensive language. Size is given in number of comments for Wiki, NC and FC domains, and number of posts for the SM domain.

Corpus	Domain	Language	Size	Type of annotation
HASOC 2019	SM	de, en, hi	24k	Hate speech, target
HatEval 2019	SM	en, es	10k	Hate speech, target, aggression
OffensEval 2019	SM	en	13k	Hate speech, target, threats
GermEval 2018	SM	de	36k	Abuse, profanity, insults
IBEREVAL 2018	SM	en,es	7k	Misogynous
MEX-A3T 2018	SM	es-mx	11k	Aggressive
Liu et al. (2018)	SM	en	30k	Hostile
Waseem 2016	SM	en	25k	Hate speech, category
de Gibert 2018	FC	en	10k	White supremacy
Gazzetta	NC	gr	1.5M	Hate speech
SFU SOCC	NC	en	663k	Toxicity, non-constructiveness
YNACC	NC	en	522k	Insults, flames
SENSEI	NC	en	2k	Quality, tone
wiki-detox 2017	Wiki	en	115k	Personal attacks, aggression, toxicity
Zhang 2018	Wiki	en	7k	Personal attacks

[1] `http://hatebase.org/`

Benchmark performance Performance varies widely with dataset and domain. OffensEval 2019 reports maximum F1 score 0.829 on the offense classification task; for the white supremacy forum comments (de Gibert et al., 2018) classification accuracy is 0.78.

3.1.2 Spam detection

Alberto et al. (2015) provide a dataset of comments on Youtube videos classified as spam or not. Several datasets are available for short text messages in social media, see e.g. (Chen et al., 2015)'s large collection of 6 million spam tweets, and the MPI collection of Twitter accounts detected as spam accounts.

Table 2: Primary existing datasets: spam.

Corpus	Size	Avail.	Language	Domain
NSC Twitter Spam	6 million tweets	public	en	Social media
Alberto 2018	1956 comments	public	en	Online comments
MPI-SWS	41,352 accounts	research	n/a	Social media

Benchmark performance Performance varies widely with dataset and domain. Wu et al. (2018) report accuracies of up to 94.5% on account classification and 88-91% accuracy on individual texts.

3.1.3 Trolling and incitement

Detection of trolling, particularly activity with a political motivation, has attracted a lot of attention. FiveThirtyEight distribute a dataset of nearly 3 million tweets sent from Twitter accounts "connected to the Internet Research Agency, a Russian "troll factory" and a defendant in an indictment filed by the Justice Department in February 2018" between February 2012 and May 2018. Narayanan (2018) then provides a smaller dataset from the same source, but annotated in more detail for level of aggression.

In our domain of UGC comments under news articles, Mihaylov and Nakov (2016) collected a dataset from over 2 years of articles (Jan 2013-April 2015) on the Bulgarian news site Dnevnik (`dnevnik.bg`), totalling 1,930,818 comments by 14,598 users on 34,514 articles. Troll comments were identified by a combination of observing other users' reactions, and checking identities in leaked documents; however, the dataset is not currently available publicly. Mojica (2017); Mojica de la Vega and Ng (2018) collected a similar dataset of English comments on Reddit, which is publicly available. New UGC trolling annotation may be required as the project progresses.

Table 3: Primary existing datasets: trolling and incitement.

Corpus	Size	Avail.	Language	Domain
FiveThirtyEight	2,973,371 tweets	public	en	Social media
Narayanan 2017	20,000 tweets	public	en	Social media
Mojica 2017	5,868 conversations	public	en	Reddit

Benchmark performance Mihaylov and Nakov (2016) achieve around 81% accuracy and F-score on the classification task, on a balanced dataset of news comments, using simple baseline linear classifiers. Mojica (2017) achieves 90% accuracy on his dataset for the trolling detection task, using a more complex conditional random field classifier.

3.1.4 Sentiment and opinion analysis

For news media comments, Celli et al. (2014) provide a corpus annotated for sentiment (positive/negative polarity together with target topic) as well as emotion towards other comment posts (appreciation towards a message topic). In other short text domains, 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; and more recently Mohammad et al. (2018) provide a dataset for a more specific task of detecting *intensity* of multi-class emotion and sentiment. Many other social media sentiment datasets are available; we begin with these shared tasks due to the relatively high reliability of their datasets.

Table 4: Primary existing datasets: conversation structure.

Corpus	Size	Avail.	Language	Domain
CoREa	2,900 comments	public	it	News comments
SemEval 2017 Task 4	50,000 tweets	public	en, ar	Social media
SemEval 2018 Task 1	11,288 tweets	public	en	Social media

Benchmark performance Mohammad et al. (2018) report correlations with gold-standard emotion intensity scores of up to 0.80 Pearson's r. Rosenthal et al. (2017) report sentiment classification accuracies up to 68% F-score.

3.1.5 Topic analysis

Topic datasets specifically for news media UGC are unfortunately rare. Some datasets were developed as part of the SENSEI project: Aker et al. (2016a) discuss a dataset with a form of automatic topic labelling (comments quoting article sentences were taken to be on-topic for that article; then randomly drawn article-comment pairs were taken as noisy off-topic instances); it also included a small test set of 100 comments from each of 18 articles, with topic labels manually annotated. However, it is not currently available. Llewellyn et al. (2014) also built a corpus of comments with topic labels, but it is small (136 and 161 comments from two articles) and does not appear to be available. Emmery (2014) collected datasets of news article comments, but provides no public dataset or annotated labels.

Datasets in similar online text are more common, but very few provide explicit manually annotated topic labels for evaluation purposes. Most use a form of *distant supervision*, assuming that topic or topical relevance can be inferred from associated metadata: the name of the online forum; the article being commented on; the question being discussed. Given this lack of availability, we propose to use the same approach, and generate new topic labels for evaluation only where required (e.g. as part of the summarisation and reporting activity T3.3). For finer-grained notions of topic, we will use the conversation structure datasets discussed in the next section, in which individual messages are labelled with particular types of contribution or with relevance relations to particular questions being discussed.

3.1.6 Conversation structure analysis

Many dialogue corpora are available that provide annotated information about conversation structure, but the majority are not directly relevant to the EMBEDDIA project purposes as they generally (a) contain only 2 speakers (making the task of identifying addressee trivial, and antecedent utterance significantly

simpler) and (b) are transcriptions or recordings of spoken language, which has very different characteristics to the written interaction of interest here. Some multi-party datasets exist that are more suitable to our purposes, and those 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. Some similar corpora with less informative annotation also exist: Bhatia et al. (2012) provide datasets from online discussion forums with messages annotated with dialogue acts including positive and negative feedback; and Catherine et al. (2012) with question-answer relations. Taking a slightly different perspective, Elsner and 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, it shares the basic problem of distinguishing conversational relevance relations between messages.

We are only aware of one dataset that contains similar 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: over 2,900 posts by 1,600 authors, containing 135,600 words).

Table 5: Primary existing datasets: conversation structure.

Corpus	Size	Avail.	Language	Domain
CoREa	2,900 comments	public	it	News comments
ICSI	75 meetings	public	en	Spoken dialogue
AMI	100 hours	public	en	Spoken dialogue
AAWD	500 threads	public	en, ru, zh	Wikipedia discussion
AACD	12 threads	public	en, ru, zh	Chat
IAC/ARGUE	11,000 threads	public	en	Online forums
Elsner and Charniak (2011)	2,601 chats	public	en	IRC chat

Benchmark performance Celli et al. (2016b) achieve c.70% F1-score for the agreement/disagreement classification problem, with agreement (F1 72.6%) better than disagreement (F1 68.4%).

3.1.7 Summarisation and reporting

Few datasets specifically for news media UGC summarisation exist; the primary resource in this domain comes from the SENSEI project (Riccardi et al., 2016), which produced an annotated corpus of human summaries of online news reader comments (Barker et al., 2016). This provides gold-standard annotations for individual comment topic labels and reference conversation thread summaries, for the comments on 18 news articles (total 1,845 comments, 87,559 words). The dataset language is English, with content sourced from the Guardian newspaper.

Table 6: Primary existing datasets: summarisation and reporting.

Corpus	Size	Avail.	Language	Location
SENSEI	1,872	public	en	`nlp.shef.ac.uk/sensei`

Our proposed approach for Task T3.3 is therefore to develop comment summarisation models for English, using both extractive approaches trained on the SENSEI dataset and other non-news UGC datasets as appropriate, and abstractive approaches based on methods developed in WP5, and apply our cross-lingual transfer approach to produce models for the project languages. Evaluation will require further annotation of the new EMBEDDIA project datasets (Section 3.2).

3.2 New datasets

In this section we describe the new collections being produced by the EMBEDDIA media partners and annotated for use in project tasks: large datasets of user comments from STY (Styria Media Servisi, Croatia) and ExM (Ekspress Meedia, Estonia). (The remaining media partner, STT do not deal with user-generated content). Given their nature as new resources, we expect the data they contain, and the metadata and annotation associated with them, to evolve considerably over the course of the project.

3.2.1 Styria user comments

Overview This dataset consists of about 30M user-generated comments from online news media sites owned and operated by Styria, and provided to EMBEDDIA by the partner Trikoder d.o.o. (previously Styria Media Services, STY).

On Styria's systems, users must be registered to post comments below news articles. Human moderators ensure that they adhere to a set of rules. Breaking the rules leads to a warning, and this fact is recorded and forms the main annotation for this dataset. As shown in the detailed description below, many rules correspond to analysis tasks from Section 2 (e.g. Rule 1 for spam detection; Rules 2,3,4,6,8 for various categories of offensive language; Rule 5 for trolling detection), although some do not (e.g. Rule 7). Other metadata include author ID, timestamp and the ID of the comment being replied to (if any); these provide annotation for conversation structure analysis tasks.

There are two types of rules - describing less serious offences (minor warnings) and more serious offences (major warnings). Breaking the rules multiple times can bring them a temporary ban or in some cases, on a discretion of a moderator, a permanent ban. User may receive up to two minor warnings - the third one leads to the short temporary ban (one day). User may receive one major warning - the second one leads to a temporary ban from the site on the period of five days. Minor and major warnings are not combined in any way - i.e. user that has two minor warnings and zero major warnings, and then receives a major warning, is not temporarily banned from the site. After the ban, the number of warnings for this type are reset.

Detailed description

- **Sources:** 24Sata (`24sata.hr`), Croatia's highest-circulation daily newspaper; Večernji List (`vecernji.hr`):
  - **24Sata:** 21.5M comments (21,548,192), all available comments September 2007-April 2019
  - **Večernji List:** 9.6M comments (9,646,634), all available comments September 2009-May 2019
- **Date of creation:** version 002, 21st June 2019.
- **Language:** Croatian
- **File format:** CSV, see Table 7 for details.

- **Annotation:** moderators' blocking decisions with reasons; see Table 8 for details.
- **Public release:** details under discussion, with intention to release 5-10% of the dataset publicly.

Table 7: Styria comment dataset file format

Variable	Variable type	Description
comment_id	integer	The internal id of the comment. Unique for each row.
user_id	string	The uuid of the user writing the comment. Unique for each user.
content	string	The content (text) of the user comment.
site	string	The site the comment came from.
reply_to_id	integer	The 'comment_id' of the parent comment - if this comment was intended as a reply.
created_date	string	The date the comment was created.
last_change	string	The date the comment was last edited.
article_id	integer	A public id of the article where this comment was posted.
infringed_on_rule	integer	If the user has infringed on rules with this comment, id of the rule is given.
like_counts	integer	A number of times other users have voted in favor of this comment, similar to the Like button.
dislike_counts	integer	A number of times other users have voted against this comment, opposite of the Like button.

Table 8: Styria comment dataset annotation schema: moderation rules

Rule ID	Warning type	Description in English (translated from Croatian original)
1	minor warning	Advertising, content unrelated to the topic, spam, copyright infringement, citation of abusive comments or any other comments that are not allowed on the portal
2	major warning	Direct threats to other users, journalists, admins or subjects of articles, which may also result in criminal prosecution
3	major warning	Verbal abuse, derogation and verbal attack based on national, racial, sexual or religious affiliation, hate speech and incitement
4	major warning	Collecting and publishing personal information, uploading, distributing or publishing pornographic, obscene, immoral or illegal content and using a vulgar or offensive nickname that contains the name and surname of others
5	minor warning	Publishing false information for the purpose of deception or slander, and "trolling" - deliberately provoking other commentators
6	minor warning	Use of bad language, unless they are used as a stylistic expression, or are not addressed directly to someone
7	minor warning	Writing in other language besides the Croatian, in other scripts besides Latin or writing with all caps
8	minor warning	Verbally abusing of other users and their comments, article authors, and direct or indirect article subjects, calling the admins out or arguing with them in any way

3.2.2 Ekspress user comments

Overview This dataset consists of about 30M user-generated comments from online news media sites owned and operated by Ekspress Meedia (ExM), and provided to EMBEDDIA by ExM.

On the Ekspress systems, users may post comments below news articles anonymously without registering. Human moderators decide whether the comments should be blocked from being published, and this fact is recorded and forms the main annotation for this dataset, for use in comment filtering and analysis tasks. Other metadata includes author ID, timestamp and the ID of the comment being replied to (if any); these provide annotation for conversation structure analysis tasks.

Detailed description

- **Source:** ExM news publications including Eesti Ekspress (`ekspress.ee`), Estonia's highest-circulation daily newspaper:
  - 31.5M comments (31,473,732), all available comments 2009-May 2019
- **Date of creation:** version 001, 7th June 2019.
- **Languages:** Estonian, Russian.
- **File format:** CSV, see Table 9 for details.
- **Annotation:** moderators' blocking decisions.
- **Public release:** TBD.

Table 9: Ekspress Meedia comment dataset file format

Variable	Variable type	Description
comment_id	string	ID of the comment
article_id	string	ID of the article for which the comment was written
created_time	string	comments creation/publish time
create_user_id	string	user ID
subject	string	title of the comment
content	string	content of the comment
replyto_comment_id	string	the parent comment ID, or None
is_anonymous	string	1 if the comment was published anonymously; 0 if the comment was published by a registered user
is_enabled	string	1 if the comment was published (online); 0 if it wasn't published
channel_language	string	language of the channel: 'nat' for Estonian, 'rus' for Russian
moderated_by	string	ID of the moderator

4 Evaluation metrics

In this section we describe metrics commonly used to evaluate classification tasks of the kind considered here, including those used in the public shared tasks used to provide existing resources.

4.1 Metrics for classification tasks

Most tasks discussed above are essentially machine learning classification tasks; therefore, we will use standard evaluation methodology and metrics from the text classification literature.

We will split the evaluation data into two sets, training set and testing set, and estimate the predictive accuracy of models on the testing set. For small datasets where a held-out set would significantly reduce the learning capability due to wasted training data, we will use cross-validation approach.

In a binary classification problem, let E denote the set of all training instances, P denote the set of positive instances, and N the set of negative instances, where $P \cup N = E$ and $|P| + |N| = |E|$. Let $TP \in P$ (true positives) be a set of positive instances that are correctly classified by the learned model, $TN \in N$ (true negatives) be a set of correctly classified negative instances, $FP \in N$ (false positives) be a set of negative instances that are incorrectly classified as positives by the learned model, and $FN \in P$ (false negatives) be a set of positive instances incorrectly classified as negative instances.

Typical metrics used in text classification are:

Classification accuracy. Classification quality of the learned models is measured by the classification accuracy that is defined as the percentage of the total number of correctly classified examples in all classes relative to the total number of tested examples. In case of binary classification problem, the accuracy of a model is computed as

$$Accuracy = \frac{|TP| + |TN|}{|E|}$$

Note that the accuracy measures the classification accuracy of the model on both positive and negative examples of the target class of interest. Instead of accuracy, results are often presented with *classification error*, which is

$$Error(Model) = 1 - Accuracy(Model)$$

Precision, recall, and F-measure In binary classification, precision is the fraction of correctly classified positive instances among all predicted as positives, i.e.

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

while recall (also known as sensitivity) is the fraction of positive instances over the total amount of positive instances.

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

A measure that combines precision and recall in a harmonic mean of precision and recall is called F_1 measure or balanced F-score:

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

4.2 Metrics for regression tasks

Some tasks above (e.g. emotion detection) can alternatively be framed as regression tasks, in which the target is not a category or class label, but a number on a continuous scale. In these cases, evaluation cannot be performed in terms of class accuracy, but is usually carried out in terms of either error or correlation. Typical metrics used in regression tasks are:

Root mean squared error. The root-mean-squared error RMSE (or root-mean-squared deviation RMSD) is an overall calculation of the amount by which a predicted distribution differs from the ideal. It is calculated by summing the squares of the differences between predicted and ideal values, normalising by the number of samples, and taking the square root:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

where y_i is the true (ideal/correct) value at sample i, and $\hat{y}_i$ the corresponding predicted value.

Pearson's correlation. The *Pearson correlation coefficient* or *Pearson's r* can be used to measure the linear correlation between the predicted and ideal distributions. It is defined as the covariance of the two variables divided by the product of their standard deviations, and thus is an appropriate measure when the relative changes are of interest, but the absolute values and absolute magnitudes of change are not.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman's correlation. The *Spearman's rank correlation coefficient* or *Spearman's rho* is a nonparametric measure of correlation, similar to Pearson's *r* in intuition but defined over the ranks of the variables rather than their values. It is therefore appropriate for cases where the relative ordering of the values is of interest, but the size of differences between them is not. It is calculated using the Pearson's *r* calculation above, after substituting variable values with their ranks in their distributions (1, 2, 3, etc).

4.3 Metrics for other tasks

Some of the analysis tasks outlined in Section 2 are not pure classification tasks, and therefore require different approaches to evaluation metrics. The two main cases are topic analysis and summarisation, and we briefly discuss these here.

4.3.1 Topic modelling

Topic modelling and assignment is notoriously hard to evaluate quantitatively, with many experiments relying on measures of instrinsic goodness of fit rather than extrinsic quality (see Blei, 2012, for discussion). Much topic modelling research is therefore evaluated qualitatively and/or by using human judges to rate or rank the quality of the topics inferred. However, where topic labels are known, standard classification metrics can be used (see above). In cases where comments must first be clustered according to topic, this is more complex as it requires soft, cluster-sensitive variants of these metrics; suitable variants have been defined (e.g. fuzzy BCubed Precision, Recall and F-Measure metrics Hüllermeier et al. (2012); Jurgens and Klapaftis (2013)) and have been used for news UGC (Aker et al., 2016a).

4.3.2 Summarisation

Summarisation is another area where evaluation can be problematic. Automatically calculated metrics such as ROUGE (Lin, 2004) exist, are widely used, and are useful for quick reference during development, but do not give a reliable measure of summary quality (see e.g. Schluter, 2017, for discussion). Research which requires more insightful and task-related evaluation therefore often relies on human judgements of summary quality.

5 Conclusions and further work

We have presented resources collected so far in order to build and evaluate technologies for analysis of user-generated content in WP3 of the EMBEDDIA project. Our repository consists of two distinct types of data: resources being generated as part of the EMBEDDIA project itself by the news media

partners, and resources collected from existing project-external sources. We expect the external collections to be used for initial development of methods and techniques, and the new collections to be used for subsequent development, cross-lingual transfer and evaluation of the final EMBEDDIA technologies. Nevertheless, we expect that the collection of datasets, and the annotation of the datasets collected, will continue throughout the course of the project: as tools to approach initial tasks (e.g., detecting comments to be moderated, or detecting sentiment) are developed, our attention will turn to finer-grained or subsidiary tasks (e.g., categorising comment types or reasons for moderation, or detecting constructive language beyond simple sentiment polarity), and these will require more detailed annotation and further evaluation, and may also lead us to collect additional semantic resources and/or require different evaluation metrics.

References

Aiyar, S. and Shetty, N. P. (2018). N-gram assisted youtube spam comment detection. *Procedia Computer Science*, 132:174 – 182. International Conference on Computational Intelligence and Data Science.

Aker, A., Kurtic, E., Balamurali, A., Paramita, M., Barker, E., Hepple, M., and Gaizauskas, R. (2016a). A graph-based approach to topic clustering for online comments to news. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR)*.

Aker, A., Paramita, M., Kurtic, E., Funk, A., Barker, E., Hepple, M., , and Gaizauskas, R. (2016b). Automatic label generation for news comment clusters. In *Proceedings of the 9th International Natural Language Generation Conference (INLG)*.

Alberto, T., Lochter, J., and Almeida, T. (2015). TubeSpam: Comment spam filtering on YouTube. In *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15)*, pages 1–6.

Badawy, A., Ferrara, E., and Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265.

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

Barker, E., Paramita, M. L., Aker, A., Kurtic, E., Hepple, M., and Gaizauskas, R. (2016). The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–52, Los Angeles. Association for Computational Linguistics.

Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B., and 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*, pages 48–57. Association for Computational Linguistics.

Bhatia, S., Biyani, P., and Mitra, P. (2012). Classifying user messages for managing web forum data. In *Proceedings of the 15th International Workshop on the Web and Databases (WebDB)*, pages 13–18.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Caruana, G. and Li, M. (2012). A survey of emerging approaches to spam filtering. *ACM Computing Surveys*, 44(2):9:1–9:27.

Catherine, R., Singh, A., Gangadharaiah, R., Raghu, D., and Visweswariah, K. (2012). Does similarity matter? the case of answer extraction from technical discussion forums. In *Proceedings of COLING 2012: Posters*, pages 175–184, Mumbai, India. The COLING 2012 Organizing Committee.

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

Celli, F., Stepanov, E., Poesio, M., and Riccardi, G. (2016a). 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)*, pages 110–118, Osaka, Japan. The COLING 2016 Organizing Committee.

Celli, F., Stepanov, E. A., and Riccardi, G. (2016b). 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*.

Chen, C., Zhang, J., Chen, X., Xiang, Y., and Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely twitter spam detection. In *2015 IEEE International Conference on Communications (ICC)*, pages 7065–7070.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

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

Elsner, M. and Charniak, E. (2011). Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189, Portland, Oregon, USA. Association for Computational Linguistics.

Emmery, C. (2014). Topic modelling in online discussions. Master's thesis, Tilburg University.

Hüllermeier, E., Rifqi, M., Henzgen, S., and Senge, R. (2012). Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Transactions on Fuzzy Systems*, 20(3):546–556.

Jurgens, D. and Klapaftis, I. (2013). Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 290–299.

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

Kantchelian, A., Ma, J., Huang, L., Afroz, S., Joseph, A. D., and Tygar, J. D. (2012). Robust detection of comment spam using entropy rate. In *Proceedings of the 5th ACM Workshop on Artificial Intelligence and Security*, pages 59–70.

Kim, D., Graham, T., Wan, Z., and Rizoiu, M. (2019). Tracking the digital traces of Russian trolls: Distinguishing the roles and strategy of trolls on Twitter. *CoRR*, abs/1901.05228.

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

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, P., Guberman, J., Hemphill, L., and Culotta, A. (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth International AAAI Conference on Web and Social Media*.

Ljubešić, N., Fišer, D., and Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in slovene and english. *CoRR*, abs/1906.02045.

Llewellyn, C., Grover, C., and Oberlander, J. (2014). Summarizing newspaper comments. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Mihaylov, T. and Nakov, P. (2016). Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany. Association for Computational Linguistics.

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

Mojica, L. G. (2017). A trolling hierarchy in social media and A conditional random field for trolling detection. *CoRR*, abs/1704.02385.

Mojica de la Vega, L. G. and Ng, V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

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

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

Narayanan, A. (2018). Tweets dataset for detection of cyber-trolls.

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

Purver, M. (2011). Topic segmentation. In Tur, G. and de Mori, R., editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317. Wiley.

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

Riccardi, G., Bechet, F., Danieli, M., Favre, B., Gaizauskas, R., Kruschwitz, U., and Poesio, M. (2016). The SENSEI project: Making sense of human conversations. In Quesada, J. and et al., editors, *Future and Emerging Trends in Language Technology*, number 9577, pages 10–33.

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

Schluter, N. (2017). The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

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

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

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Wu, T., Wen, S., Xiang, Y., and Zhou, W. (2018). Twitter spam detection: Survey of new approaches and comparative study. *Computers and Security*, 76:265–284.