Incendiary News Detection

Enis Berk Coban, Elena Filatova
City University of New York
ecoban@gradcenter.cuny.edu, efilatova@citytech.cuny.edu

Abstract

We introduce the problem of *incendiary news detection*. We compare and contrast this problem with the problem of hate speech detection in social media. Most of the social media posts that are classified as hate speech contain straightforward slurs, insults, swearing, etc. In contrast to social media posts, incendiary news articles often do not contain any straightforward slurs and insults but, nevertheless, incite hate. To detect such news articles, we leverage a resource from the Turkish community, where activists attempt to combat hate in media by manually tagging the news articles inciting hate. To collect non-incendiary news we retrieve news articles from two news agencies: BBC and CNN. Both BBC and CNN are recognized worldwide as serious media and are unlikely to contain foul language. We collect three different non-incendiary news corpora to ensure the validity of our classification results. We use several feature sets and classification approaches to differentiate between incendiary and non-incendiary news. Our classification system achieves 97.0% accuracy.

1 Introduction

There exist several systems whose goal is to detect hate speech in social media, e.g., Twitter messages, Facebook posts, various discussion forums, etc. (Tulkens et al. 2016; Schmidt and Wiegand 2017; Fortuna and Nunes 2018). As the emphasis is on social media posts, the object under analysis is typically a short text message with a direct addressee of hate speech. The most successful classification features used for hate speech detection are word and character n-grams, and abusive language lexicons. These features capture the language that hate-speech messages on social media tend to use.

In this project, we study the news articles that ignite hatred. We call such news articles *incendiary news*. In modern society, it is a common for some news outlets to publish different opinions on the same event, person, issue. However, when an article's goal is to incite hatred, this is hardly a "difference of opinions." In contrast to social media messages, though, such news articles are typically edited, "groomed" and, thus, do not contain slur words, direct insults, or other straightforward examples of foul language that targets a particular addressee. According to van Dijk (2006), "*International research of the last three decades has consistently found that the European news media in general, and the written press in particular, have been part of the problem of racism, rather than part of its solution.*" We believe, this finding can be generalized to the whole world (not just Europe) and to many issues (racism being only one of them).

In Turkey, a non-profit organization, Hrant Dink foundation, runs a project called *Nefret Söylemi (hate speech)*.[1] The foundation was created in memory of Hrant Dink, a journalist who is best known for advocating Turkish-Armenian reconciliation and supporting human rights movement in Turkey. The main objective of the *Nefret Söylemi* project is "to combat racism and discrimination based on ethnic and religious grounds, through monitoring the newspapers and exposing the *problematic* articles in the media." The project participants monitor news written in the Turkish language and manually annotate incendiary news.

Collecting and tagging incendiary news articles manually is a time consuming task. In this work, we utilize the enormous work already completed by the Hrant Dink Foundation: we download the news articles that have been collected and labeled as incendiary news. To collect examples of non-incendiary news articles that describe the same events, people, issues that are used in the manually labelled collection of incendiary news, we use the Turkish language BBC[2] and CNN[3] web sites. Irrespectively of the BBC's and CNN's political alliances, we rely on BBC's and CNN's reputation and assume that the news articles published on either BBC or CNN do not have any hate-inciting content.

We want to emphasize that we use three different non-incendiary news corpora. It is expected that cross-corpus training for text classification might lead to lower classification accuracy. Rangel et al. (2018) demonstrate that among the reasons for lower cross-corpus classification results for the task of Native Language Identification is the fact that the language model captures the differences in the described topics rather then the peculiarities attributed to the speakers

[1] http://www.nefretsoylemi.org/

[2] https://www.bbc.com/turkce

[3] https://www.cnnturk.com/

of various languages. To avoid this issue we (1) collect initial BBC and CNN corpora using the news topics covered in the incendiary news; (2) use the collected BBC corpus and the manually annotated corpus of incendiary news to obtain the terms that are most descriptive of the incendiary news; (3) use these terms to collect another BBC corpus making it as close as possible to the set of manually annotated incendiary news with regard to the information coverage.

The contribution of our work is three-fold:

- We extend the problem of hate speech identification and introduce the problem of incendiary news detection.
- We propose a novel approach for corpus generation and use this approach to collect three different sets of non-incendiary news to ensure that both incendiary and non-incendiary news cover the same topics, issues, people, events, etc. All the corpora of incendiary and non-incendiary news that are used in this project are made available for research purposes.[4]
- We run a set of classification experiments and demonstrate high accuracy results in distinguishing between incendiary and non-incendiary news.

The rest of the paper is organized as follows. In Section 2, we describe the related work. In Section 3, we discuss the process used to collect incendiary news; as well as the process that we suggest to collect three sets of non-incendiary news. We undertake the effort of collecting three different non-incendiary news collections to ensure that while creating a classification model we, indeed, capture the difference between incendiary and non-incendiary news rather than the difference in the list of topics covered in incendiary vs. non-incendiary news articles. In Section 4, we discuss the classification experiments and our unique approach for creating classification model by using different corpora of non-incendiary news on different stages of classification. Finally, in Section 5 we outline avenues for our future research.

2 Related work

Hate speech detection has been a topic of interest for the Text Mining community before the proliferation of social media. Spertus (1997) uses rule-based decision tree classification to identify hostile messages on web forums.

Greevy and Smeaton (2004) use SVM with bag-of-word (BOW) and bi-gram features to successfully differentiate between racist and non-racist web pages. The best results (92.78% precision and 90% recall) are obtained using BOW features.[5]

Currently, many researchers work on identifying hate speech in social media posts. The social media hate speech detection systems work with: Twitter (Waseem 2016; Park and Fung 2017), user comments to news articles (Nobata et al. 2016), Wikipedia discussions (Wulczyn, Thain, and Dixon 2017), Facebook posts (Del Vigna et al. 2017), Instagram comments (Zhong et al. 2012).

BOW and n-grams are among the most reliable features for detecting hate speech in social media posts (Chen et al. 2012; Razavi et al. 2010; Warner and Hirschberg 2012; Nobata et al. 2016). Another source of features strongly associated with hate speech are lexicons of abusive language (Spertus 1997; Razavi et al. 2010; Gitari et al. 2015).

In addition to word n-grams and lexicons that deal with the text on the word or phrase level, hate speech detection systems successfully employ character n-grams (Mehdad and Tetreault 2016). The success of character n-grams is due to the fact that special characters are frequently used to mask slurs. This technique is now popular as many social media platforms prohibit the use of abusive words.

Thus, the most successful classification features that can be reliably used to identify hate speech in social media (lexicons, word n-grams, character n-grams) capture the straightforward use of inappropriate language.

Following the assumption that hate speech is often associated with negative sentiment, several hate detection systems employ sentiment detection as part of the hate speech detection process (Dinakar et al. 2012; Sood, Churchill, and Antin 2012). Several hate speech detection projects (Nobata et al. 2016; Kshirsagar et al. 2018) take advantage of the recent advances in deep learning and use word embeddings (Le and Mikolov 2014) for hate speech detection.

More information on hate speech detection in social media posts can be found in these two surveys: (Schmidt and Wiegand 2017; Fortuna and Nunes 2018).

Not only social media posts have content that can be classified as hate speech. An information retrieval system presented by Greevy et al. (2004) deals with the problem of identifying hate (e.g., racism) on web pages. Identifying web pages that incite hatred is an important part of marketing research (Fan and Chang 2010) as most companies do not want to be associated with hate speech and violent content (Solon 2017). Journalists community strives to maintain high journalism standards (van Dijk 2006).

In this work we deal with the problem of incendiary news detection which we compare to the problem of hate speech detection in social media posts.

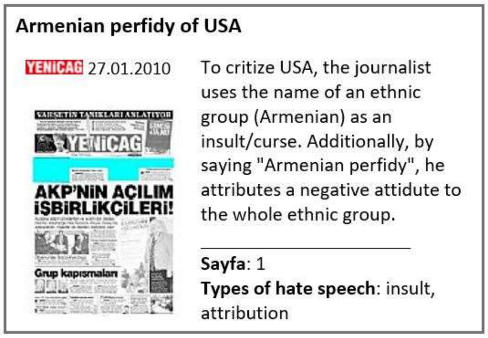
3 Corpus Collection

Corpus collection for the task of hate speech or incendiary new detection is extremely hard. Currently, many Twitter generated corpora use the author injected hashtags for automatic corpus generation (Muresan et al. 2016). However, it is highly unlikely that an author of a twitter hate speech message would mark this message with the *#hatespeech* or similar hashtags. Thus, many existing hate speech corpora are generated using a set of curse words from user generated dictionaries (Xiang et al. 2012; Kwok and Wang 2013; Raisi and Huang 2016). Another approach for corpus generation frequently used by the NLP comminuty relies on crowdsourcing platforms (Davidson et al. 2017).

Given the severity of the hate speech problem in the modern world, there are organizations, projects, activists who search for hate-infused messages and incendiary news, label and report them. In our work, we use the data labeled by

[4] https://github.com/EnisBerk/Incendiary_news

[5] The corpus used by Greevy and Smeaton (2004) is currently unavailable.



Armenian perfidy of USA

YENİCAG 27.01.2010

To critize USA, the journalist uses the name of an ethnic group (Armenian) as an insult/curse. Additionally, by saying "Armenian perfidy", he attributes a negative attidute to the whole ethnic group.

Sayfa: 1

Types of hate speech: insult, attribution

Figure 1: An example of a news article identified as incendiary by *Nefret Söylemi*.

one of such watchdog organizations. We hope that the system created as part of this research project can be used by this organization to faster locate incendiary news.

In the absence of blatant expression of hate, differentiating incendiary from non-incendiary news is hard, as it relies on subtle nuances and cultural context. Like with many types of figurative language (e.g., sarcasm, metaphor), the meaning on the text goes beyond the surface form. This issue becomes even more pronounced when dealing with text that avoids the explicit usage of slurs or offensive terms to convey hateful messages.

In this paper, instead of trying to devise our own definition of incendiary news, we adopt the annotation strategy of an NGO watchdog organization who developed their own criteria for monitoring and flagging news as incendiary. In future we plan to explore the notion of incendiary news in other languages to identify what makes news articles incendiary across different languages, cultures, countries.

Collecting Incendiary News

We use the data collected and labeled by the Turkish foundation named after Hrant Dink. This foundation runs a project called *Nefret Söylemi (hate speech)*. The goal of the project is to locate and annotate incendiary news articles in Turkish. The annotation includes: the picture of the article (in PDF format); the newspaper title; page number where the article was published; date the article was published; a brief (optional) abstract of the article in either English or Turkish; tags identifying why this article is an example of an incendiary article. Figure 1 is an example of such annotation. The article from in Figure 1 was published on January 27, 2010; on Page 1 (Sayfa 1) of the newspaper *Yeniçağ*.[6] According to the manual annotation, this article contains an insult and attribution. The annotation of this news article also includes its summary in English done by the *Nefret Söylemi* annotators. Figure 1 is the exact image that was posted on the *Nefret Söylemi* project web site.[7] Translation, terminology, orthography, etc. are preserved.

Keyword in Turkish	Translation into English	Inced. news	BBC-1	CNN
mülteci	refugee	51	190	240
gavur	unbeliever	36	11	2
işbirlikçi	coconspirator	31	67	0
türk düşmanı	turcophobe	17	32	240
yobaz	zealot	7	3	0
müslüman	muslim	417	190	240
yahudi	Jew	318	190	90
ermeni	Armenian	293	190	212

Table 1: Search terms used in the *Nefret Söylemi* project; the number of news containing these terms.

To locate incendiary news, the *Nefret Söylemi* project activists use a set of keywords corresponding to the topics that are sensitive or controversial in the Turkish language community. Six of these keywords are listed in the 2015 *Nefret Söylemi* project report.[8] This report also states that in 48% of the annotated news articles, the target of hate are Jews and Armenians. Table 1 lists the six keywords from the report plus two ethnicities as well as their translations into English.

Using the links to the articles listed as incendiary by the *Nefret Söylemi* project's web site, we collect 1036 news articles in Turkish (from Turkish news media). Table 1 (column *Inced. news*) contains the number of the incendiary news in our corpus (1036 news articles) that contain these keywords. Many incendiary articles contain more than one of the keywords listed in Table 1. The numbers in this table correspond to the exact matches of the keywords in text without taking into consideration the highly productive inflectional and derivational morphology of Turkish.

The *Nefret Söylemi* web site contains links to more than 1036 articles manually annotated as incendiary. However, only pictures of the newspaper pages containing the spotted articles are published on the *Nefret Söylemi* web site. Thus, automatic crawling of this site and retrieving the articles listed on this site is a complex task. To collect the set of incendiary articles from the *Nefret Söylemi* web site we use two methods. First, we use the Tesseract Open Source OCR Engine.[9] Second, we query Google using the title of the news article and the name of the newspaper where the article is published. Using the OCR Engine method we collect 80% of the incendiary news articles in our corpus. By retrieving news articles using the news titles and newspaper names as search queries we collect 20% of the incendiary news articles in our corpus. Neither of the described methods results in a clean news article text. Thus, all the documents are manually checked and cleaned to ensure the quality of the collected data.

[6] http://www.yenicaggazetesi.com.tr

[7] http://www.nefretsoylemi.org/en/detay.asp?id=91&bolum=bizden

[8] http://www.nefretsoylemi.org/rapor/Ocak-Nisan-2015-raporu_online_versiyon.pdf

[9] https://github.com/tesseract-ocr/tesseract

Collecting Non-Incendiary News: Step 1

To collect the non-incendiary part of the corpus we use the Turkish-language versions of the BBC and CNN web sites. We rely on the reputation of BBC and CNN and assume that irrespectively of the topics discussed in the articles published on these web sites, these articles are not incendiary. We understand that both BBC and CNN are not without bias but we believe that irrespectively of the topic and the opinion expressed in the news articles published by BBC and CNN, both of these sources can be used as the first approximation of non-incendiary news articles. To ensure the quality of the collected corpus we target those non-incendiary articles that discuss the same topics as the news articles that are manually annotated as incendiary. Otherwise, it is easy to come up with nearly perfect classification by simply using different news topic categories in the corpus (e.g., news articles about politics vs. news articles about sports).

Unfortunately, it is not possible to retrieve the exact counterpart non-incendiary news articles that correspond exactly to the events, people, issues discussed in the incendiary news. Thus, one of the issues that we face at the stage of corpus generation is how to ensure that the collected incendiary and non-incendiary documents cover the same set of topics. This issue is of paramount importance as we need to generate the corpus that can be used for creating models to identify incendiary news based on the presence of hatred in these documents rather than based on the set of topics covered in these documents. To address this issue we introduce a novel two-step procedure for non-incendiary news collection. All in all, we generate three different sets of non-incendiary news, one of which is used to fine tune the non-incendiary news collection procedure, and the other two are used in the classification experiments.

On Step 1, the best we can do is to search BBC and CNN web sites using the keywords that correspond to sensitive topics and are used to locate the incendiary news. Thus, to retrieve the non-incendiary part of our corpus we use the BBC and CNN web sites' search bars and submit as queries the keywords listed in Table 1. This way we collect two sets of non-incendiary news: CNN (948 docs) and BBC-1 (1031 docs). Table 1 contains the terms used to search for non-incendiary news in BBC-1 and CNN and the number of the documents retrieved from BBC and CNN using these keywords.

Collecting Non-Incendiary News: Step 2

It must be pointed out that the collected CNN and BBC-1 corpora are created using rather broad keywords as search queries. Out goal, however, is to collect the non-incendiary news articles that are as close as possible to the incendiary news articles in terms of the topics, issues, events covered.

Thus, we use the BBC-1 data set to collect another set of non-incendiary news. To do this we run a classification experiment that allows us to identify those keywords that are typical for incendiary news articles. In this classification experiment we use the set of the collected incendiary news and the BBC-1 set of non-incendiary news. For the moment, we are not interested in the classification results, rather, we are interested in the information gain for the words as classification features. We use information gain to identify which words are the best predictors of incendiary versus non-incendiary news.

Thus, we identify 20 Turkish terms that are typical for the incendiary news in our corpus. These 20 Turkish terms (with the corresponding English translations) are: *Ak* (white), *Allah* (God), *aşırı* (excessive), *barış* (peace), *CHP* (abbreviation for the main opposition party), *cumhurbaşkanı* (president of the republic), *din* (religion), *dışişleri* (foreign affairs), *Ermeniler* (Armenians), *gün* (day), *haçlı* (crusader), *İslam* (Islam), *kadın* (woman), *kardeşler* (brothers/sisters), *Müslüman Kardeşler* (Muslim Brothers), *Mescid-i* (Mosque of), *Milli* (national), *mülteci* (refugee), *Osmanlı* (Ottoman), *Rum* (Greek), *savaşa* (to war), *Suriye* (Syria), *Suriyeli* (Syrian), *sözde* (so-called), *terör* (terror), *Türk* (Turk), *Türkiye'nin* (Turkey's), *üniversitesi* (university of), *şehit* (martyr). We then use these 20 terms to collect 50+ documents for each of these terms using the BBC's web site search tool bar. Thus, we obtain the BBC-2 collection of 1038 non-incendiary news articles. To avoid overfitting we do not use BBC-1 corpus in our classification experiments.

Thus, our final corpus consists of

- 1036 manually labeled incendiary news articles;
- 1038 non-incendiary news articles retrieved from BBC (BBC-2 data set);
- 948 non-incendiary news articles retrieved from CNN.

We keep two sets of non-incendiary news in our corpus for cross-corpus training. Rangel et al. (2018) notice that for the task of native language detection, using different corpora for training and testing often leads to the significant drop in the results. They suggest that this is likely due to the fact that the classification model captures the differences between the topics described in different corpora rather than the differences in the language peculiarities that they intend to capture. We believe, this problem is relevant to the incendiary news detection task as well.

Incendiary News vs. Hate Speech in Social Media

As already mentioned, the most reliable classification features for detecting hate speech in social media posts are lexicons, word and character n-grams. These features work so well due to the fact that hate speech on social media is often associated with straightforward foul languages: abuses, insults, swearing, etc. According to our hypothesis, most incendiary news are edited, "groomed" and do not contain straightforward foul language.

There is no ready-to-use lexicon of Turkish slurs that we could use. Thus, to test our hypothesis, specifically for this work, we create a list of frequent 40 Turkish slurs. We extend this list with the morphological variations of these 40 slurs. We then check the presence of these words in the incendiary news subset of our corpus. Only 9% of the incendiary news articles in our corpus contain the words from our lexicon. The problem of lexicon generation is beyond the scope of this paper, but we believe that this small experiment allows us to say that the surface manifestation of hate in news articles differs from the one on social media.

4 Experiments

The goal of our classification experiment is to create a model that differentiates between incendiary and non-incendiary news. The preprocessing and feature selection is done using the NLTK.[10] For the classification experiments we use the SciKit-Learn.[11] For stemming we use TurkishStemmer.[12]

Data

For our experiments we use the corpus of news articles in Turkish described in Section 3. Our corpus consists out of three parts: incendiary news articles and two sets of non-incendiary news articles. The idea of using three-part corpus for binary classification comes from the idea of cross-corpora training. In our case, we have only one set of the incendiary news articles and two sets of non-incendiary news.

Before starting the classification experiment we preprocess the data:

- remove URLs inside the articles;
- remove stop words;
- remove all non-Turkish alphabet characters (including numbers);
- remove news articles header information (e.g., newspaper's name, author's name, date).

Features and Classifiers

In our experiments we use three classification methods: Linear Support Vector Classification (SVC); binary Naïve Baysian; feedforward neural network (a multilayer perceptron). To run our experiments we use the default parameters of SciKit-Learn library for these classifiers. We experiment with word and character n-grams. We also use Turkish word embeddings trained on *Common Crawl*[13] and *Wikipedia* using *fastText* (Grave et al. 2018).

Two Classification Experiments

To ensure the quality of our conclusions, we run two sets of classification experiments: (1) using the BBC-2 corpus for both training and testing; (2) using the BBC-2 corpus for training and CNN corpus for testing. In both cases, we use 80% of the collected incendiary news articles for training and 20% of the collected incendiary news articles to testing. For the first experiment we use 80% of the BBC-2 non-incendiary news articles for training and 20% of the BBC-2 non-incendiary news articles for testing. For the second experiment we use the model obtained for the first experiment and test this model using 20% of randomly selected CNN non-incendiary news articles.

Results

Table 2 contains the results of the classification experiments using word n-grams, character n-grams, and word embeddings as features. We run classification experiments for unigrams, and combination of uni- and bi-grams.

For the cross-corpus classification experiment, there is an expected overall drop in the performance as compared to the within corpus classification. However, for several classification features-algorithm pairs, this drop is not dramatic. This allows us to conclude the that the BBC and CNN language models are very similar. We believe that this proves that the corpus generation procedure we suggest and apply for our corpus generation successfully captures the non-incendiary news that are close in terms of topic coverage to the manually annotated incendiary news. Thus, the obtained language models capture the differences among incendiary and non-incendiary news.

When word embeddings are used as features, the results are slightly lower than the results obtained using word n-grams. We believe, it is due to two reasons: (1) our corpus is rather small and word embeddings obtained using Turkish Wikipedia overgeneralize the language model; (2) the embeddings that we use are obtained from Turkish Wikipedia whose language might differ from the language used in news articles.

The results discussed above are obtained without stemming. We run addition experiments with stemming and obtain similar results. We omit these results in this paper due to the FLAIRS paper length restrictions.

As discussed in Section 3, in contrast to hate speech present in social media, incendiary news do not contain straightforward foul language. However, our experimental results (Table 2) demonstrate that the difference in the choice of words for incendiary and non-incendiary news is substantial and allows to create a reliable model that differentiates between incendiary and non-incendiary news.

5 Future Work

The results presented in Table 2 demonstrate that most standard classifiers with standard feature sets (word and character n-grams) can reliably distinguish between incendiary and non-incendiary news articles in Turkish.

In future, we are interested in two research avenues: analyzing cross-lingual and cross-cultural aspects on incendiary news; and in studying the existing style transfer efforts with the goal of applying their finding to automatically re-write incendiary news to eliminate hatred from them.

We also hope that the system developed as the result of this project can be used by NGO organizations to help their activists to monitor the news landscape, search and report incendiary news.

References

Chen, Y.; Zhou, Y.; Zhu, S.; and Xu, H. 2012. Detecting offensive language in social media to protect adolescent online safety. In *PASSAT/SocialCom*.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *ITASEC*.

[10] https://www.nltk.org/

[11] http://scikit-learn.org

[12] https://github.com/otuncelli/turkish-stemmer-python

[13] http://commoncrawl.org/

		Within Corpus (BBC-2: train & test)			Cross-Corpus (BBC-2: train; CNN: test)		
Classifier	**Feature** (**character** n-grams)	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Linear SVC	2-gram	0.93	0.95	0.94	0.74	0.95	0.83
Naïve Baysian	2-gram	0.97	0.86	0.91	**0.88**	0.86	0.87
Multi-layer Perceptron (MLP)	2-gram	0.93	0.96	0.95	0.77	0.94	**0.95**
Classifier	**Feature** (**word** n-grams)	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Linear SVC	1,2-gram	**0.98**	0.97	0.97	0.77	0.97	0.86
Naïve Baysian	1,2-gram	**0.98**	0.95	0.97	0.87	0.95	0.91
Multi-layer Perceptron (MLP)	1-gram	0.96	**0.99**	**0.98**	0.80	**0.99**	0.88
Classifier	**Feature** (word embeddings)	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Linear SVC		0.97	0.90	0.94	0.95	0.95	0.92
Naïve Baysian		0.96	0.85	0.90	0.89	0.85	0.87
Multi-layer Perceptron (MLP)		0.95	0.97	0.96	0.82	0.97	0.89

Table 2: Classification results.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM TiiC Journal*.

Fan, T.-K., and Chang, C.-H. 2010. Sentiment-oriented contextual advertising. *KAIS Journal* 23:321–344.

Fortuna, P., and Nunes, S. S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*.

Gitari, N.; Zuping, Z.; Damien, H.; and Long, J. 2015. A lexicon-based approach for hate speech detection. *Journal of Multimedia and Ubiquitous Engineering* 10:215–230.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC 2018*.

Greevy, E., and Smeaton, A. 2004. Classifying racist texts using a support vector machine. In *SIGIR*.

Kshirsagar, R.; Cukuvac, T.; McKeown, K.; and McGregor, S. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of ALW2*.

Kwok, I., and Wang, Y. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *JMLR*.

Mehdad, Y., and Tetreault, J. 2016. Do characters abuse more than words? In *SIG on Discourse and Dialogue*.

Muresan, S.; Gonzalez-Ibanez, R.; Ghosh, D.; and Wacholder, N. 2016. Identification of nonliteral language in social media: A case study on sarcasm. *JASIST* 67.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the 25th WWW conference*.

Park, J. H., and Fung, P. 2017. One-step and two-step classification for abusive language detection on Twitter. In *ALW1*.

Raisi, E., and Huang, B. 2016. Cyberbullying identification using participant-vocabulary consistency. In *#Data4Good*.

Rangel, F.; Rosso, P.; Uitdenbogerd, A. L.; and Brooke, J. 2018. Cross-corpus native language identification via statistical embedding. In *NAACL Workshop on Stylistic Variation*.

Razavi, A. H.; Inkpen, D.; Uritsky, S.; and Matwin, S. 2010. Offensive language detection using multi-level classification. In *Canadian Conf. on AI*. Springer.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Workshop on NLP for Social Media*.

Solon, O. 2017. Google's bad week: YouTube loses millions as advertising row reaches US. In *The Guardian, 03/25/17*.

Sood, S.; Churchill, E.; and Antin, J. 2012. Automatic identification of personal insults on social news sites. *JASIST*.

Spertus, E. 1997. Smokey: Automatic recognition of hostile messages. In *IAAI/AAAI*.

Tulkens, S.; Hilte, L.; Lodewyckx, E.; Verhoeven, B.; and Daelemans, W. 2016. The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands*.

van Dijk, T. A. 2006. Racism and the European Press. In *ECRI*.

Warner, W., and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In *LSM*.

Waseem, Z. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *NAACL NLP and Computational Social Science*.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *WWW*.

Xiang, G.; Fan, B.; Wang, L.; Hong, J. I.; and Rose, C. P. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *CIKM*.

Zhong, H.; Li, H.; Squicciarini, A. C.; Rajtmajer, S. M.; Griffin, C.; Miller, D. J.; and Caragea, C. 2012. Content-driven detection of cyberbullying on the Instagram social network. In *IJCAI*.