Towards the Automatic Classification of Offensive Language and Related Phenomena in German Tweets

Julian Moreno Schneider, Roland Roller, Peter Bourgonje, Stefanie Hegele, Georg Rehm
DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
Corresponding author: julian.moreno_schneider@dfki.de

Abstract

In recent years the automatic detection of abusive language, offensive language and hate speech in several different forms of online communication has received a lot of attention by the Computational Linguistics and Language Technology community. While most approaches work on English data, publications on languages other than English are rare. This paper, submitted to the GermEval 2018 Shared Task on the Identification of Offensive Language, provides the results of several experiments regarding the classification of offensive language in German language tweets.

1 Introduction

In recent years the automatic detection of abusive language, offensive language and general hate speech comments in several different forms of online communication (e. g., Twitter, Facebook, and other forms of social media or, more generally, user-generated content) has received a lot of attention by the Computational Linguistics and Language Technology community. One of the underlying assumptions of nearly all approaches published so far is the idea of setting up a watchdog service that is able to detect instances of offensive language, abusive language, hate speech, or cyberbullying, among others, fully automatically – and with high classification precision – in order to prevent the specific message or content from being posted or to flag the respective piece of content to human experts monitoring the respective system so that they can initiate corrective actions.

While most approaches towards the automatic detection of offensive online language work on and with English data sets, publications on languages other than English are rare. This article, submitted to the GermEval 2018 Shared Task on the Identification of Offensive Language, provides the results of several experiments regarding the classification of offensive language in German language tweets.

The remainder of this article is structured as follows. First, Section 2 provides an overview of related work, while Section 3 briefly describes the data set used in the GermEval 2018 Shared Task on the Identification of Offensive Language as well as the two classification tasks and their respective categories. Section 4 characterises the experiments we carried out including features and classifiers used. Section 5 briefly sketches the results of the experiments, while Section 6 lists the six runs submitted to the Shared Task. Section 7 discusses our results and Section 8 concludes the article.

2 Related Work

Recent years have seen an increasing amount of attention from the NLP community to hateful conduct and aggression online. While at first glance separating constructive, useful content from, for example, hate speech might seem like a typical text classification problem, comparable to spam classification and sentiment analysis where typical text classification approaches may be well applicable, the question whether or not certain utterances are still acceptable within the boundaries of free speech puts this task in the intersection of several research areas and disciplines, including linguistics, sociology (Jones et al., 2013; Phillips, 2015), psychology (Kowalski and Limber, 2013; Dreißing et al., 2014), law (Marwick and Miller, 2014; Banks, 2010; Massaro, 1991) and also common sense. An overview of current NLP-based approaches is collected and presented in Schmidt et al. (2017).

The complexity of the task results in a variety of difficulties that have yet to be solved. What should be considered as offensive, racist, sexist or profane, and the extra-linguistic nature of the issue are complicating factors. The nature of an utterance often depends on factors like context, (ethnicity of the) author, (ethnicity of the) targeted person or group,

Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 21, 2018

whether or not irony is the case, etc. (Nand et al., 2016; Waseem et al., 2016; Warner et al., 2012). All of this makes the creation and annotation of corpora a challenging task. Currently there is no large, universally used data set available. Numerous data sets have been created for specific tasks differing in size (from a couple of hundred labelled tweets to hundred thousands of labelled discussions) as well as text genres, e.g., Twitter (Burnap et al., 2015; Waseem, 2016; Waseem et al., 2016; Davidson, 2017), Yahoo! (Djuric et al., 2015; Nobata et al., 2016) and Wikipedia (Wulczyn et al., 2017).

Most related work on detecting abusive language has been done for English, focusing on the data set by Waseem (2016) annotated for the three categories "Sexism", "Racism" and "Other". Many approaches rely on supervised learning with Support Vector Machines as the most frequently used classifier (Davidson, 2017; Bourgonje et al., 2017). Recent approaches employing deep learning architectures have shown to compete with or even outperform these approaches. For the task on distinguishing the three categories named above the best result (F-score of 0.93) was reached by Badjatiya et al. (2017) using an LSTM model with features extracted by character n-grams, and assisted by Gradient Boosted Decision Trees. Park et al. (2017) implemented three CNN-based models for classification. Pitsilis et al. (2018) suggested a detection scheme consisting of Recurrent Neural Network (RNN) classifiers.

3 Data Set and Tasks

The GermEval 2018 task focuses on the linguistic analysis of offensive content in German tweets, 5009 of which were provided as training data.[1] A detailed description of the annotation process along with the annotation guidelines was also made available. There are two different tasks with the provided training data annotated as follows. Task 1 is a binary classification task deciding whether a tweet is offensive or not (labels OFFENSIVE: 1688, OTHER: 3321). Task 2 is a fine-grained classification task distinguishing four subcategories (labels PROFANITY: 71, INSULT: 595, ABUSE: 1022 and OTHER: 3321). The data set consists of tweets only without any kind of meta information such as the tweet ID etc. The average token size per tweet is 21,9 and consists of 1,6 sentences.

Related tasks for English such as the Workshop on Abusive Language Online (ALW)[2] have chosen different sets of data labels ranging from binary classification (e. g., PERSONAL ATTACK vs. NONE in a Wikipedia corpus (Wulczyn et al., 2017) to more granular tag sets (e. g., RACISM, SEXISM and NONE, applied to Twitter data (Waseem, 2016)). Transparent annotation guidelines are not always made publicly available, making attempts of leveraging knowledge from related data sets a formidable challenge (see the experiments on crosslingual embeddings in Section 7).

4 Experiments

We follow the majority of earlier work in this field, as described in Section 2, that employs neural networks to implement classifiers to tackle the challenge. The data and individual messages in the GermEval 2018 Shared Task is challenging due to their short length (i. e., tweets) and due to the annotated categories that are, conceptually, relatively close to one another. As reflected by rather low inter-annotator agreement scores reported for similar annotations on comparable data sets, when intellectually exploring training data, even for humans it is challenging to reliably and consistently assign labels to tweets or, on a more abstract level, to agree what constitutes "abusive" or "offensive language". In an attempt to find the best way of solving this task using a neural network approach, we not only experimented with different network architectures, but also made an effort to obtain and include additional training data as well as to enrich the given tweets with additional meta information.

4.1 Data Enrichment

Below we present the various techniques to enrich tweets by additional information as well as an automatic generation of further training data.

Gender Information Extra-linguistic information about tweets can be decisive when making a final call on whether or not some piece of content should be considered insulting, profane, abusive or non-offensive. Retrieving identity information of the author would be valuable information to classify content more reliably. Since getting this type of metadata in the form of the user ID is typically not feasible for such data sets, we attempted to classify for one aspect of user identity, i. e., the gender

[1] https://projects.fzai.h-da.de/iggsa/projekt/

[2] https://sites.google.com/site/abusivelanguageworkshop2017

Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 21, 2018

of the author. We experimented with augmenting the GermEval tweets with gender information to establish whether or not this feature would be helpful in classification. To obtain gender labels for the tweets, we scraped the tweets annotated for the TwiSty corpus (Verhoeven et al., 2016) and classified this using FastText[3] (Joulin et al., 2016), achieving an accuracy of 79.77 for this binary classification task. The GermEval tweets were then labeled using this classifier. The results using this as an additional feature in the classification of the test set are included in Tables 1 and 2.

User Profile Information As another piece of extra-linguistic information, user profile information of Twitter users mentioned in tweets were retrieved. For example, for the tweet [*@StephanJBauer @soskinderdorf Auch in Deutschland hungern Kinder.*], we retrieved the profile descriptions for *@StephanJBauer* and *@soskinderdorf* and added this to the representation of the tweet. The rationale behind this is that certain users with a particular (potentially controversial) political profile and high visibility could be more likely to trigger offensive tweets (i. e., we attempt to model the identity of the target audience, and not that of the author). The results for this setup are included in Tables 1 and 2.

Sentiment Another linguistic feature that we have included is sentiment analysis. This processing step was carried out using a simple dictionary lookup using the data set published by Waltinger (2010). According to the largest number of positive/negative sentiment words found in the tweet, we assigned the labels POSITIVE, NEGATIVE, NEUTRAL and POS_NEG in case the tweet has as many positive as negative sentiment words.

Additional User Friend Data Lastly a set of automatically labelled tweets for Task 1 is generated in order to increase the size of the data set to train the classifier. For this purpose, a small subset of the training data (70 tweets) has been selected to identify the original source (user) of the tweet. From this subset 25 different users have been identified. Most users occurred various times and in various cases it turned out that a user who posted an OFFENSE tweet might have also posted OTHER tweets. However, users who posted an OFFENSE tweet at least once were assigned to the OFFENSE user group and all others to the OTHER group. Using this list of users a set of approximately $4,000$ tweets could be automatically labelled. Thus, a tweet from a person of the OFFENSE group was automatically labelled as OFFENSE and a tweet from a person of the OTHER group as OTHER. In order to further increase the data size, the user list has been extended by taking all twitter friends into account, assigning each person of the friend list to the same user group. In this way a list of $25,000$ users has been created, resulting in 2 million automatically labelled tweets. In order to stick to a practical and feasible setup, i. e., to be able to run the experiments on standard hardware, automatically labelled data was reduced to $50,000$ tweets using the same ratio of OTHER/OFFENSE as in the manually labelled training data. This set of tweets is not added to each tweet as a feature, but used as a new training corpus, i. e., the neural network is first trained with the new corpus of automatically obtained tweets and then the training is refined with the training set of GermEval 2018.

4.2 Architecture

To set a baseline performance we use FastText, which allows for both supervised and unsupervised text classification combining word embeddings with character n-grams instead of CBOW (which is the case for Word2vec). We apply out-of-the-box supervised classification using Wikipedia embeddings to obtain our baseline score. In addition to that we generate embeddings from a German Twitter snapshot described by Scheffler (2014). Due to its higher flexibility we use Keras[4] for all other experiments reported on in this paper.

The neural network that we implemented and tested is based on the architecture by Wang et al. (2017). Their architecture is composed of three layers: (i) a convolutional layer; (ii) a MaxPooling layer; and (iii) a dense layer, that performs the classification itself. We made minor modifications to this setup and instead of using one convolutional layer and one dense layer, we use two convolutional and two dense layers. Due to the relatively large number of dimensions (300), any relevant information in the input data would be better preserved with two convolutional layers. The second dense layer is there to accommodate the more detailed classification for Task 2, which not only comprises more classes but also classes that are conceptually

[3] https://fasttext.cc

[4] https://keras.io

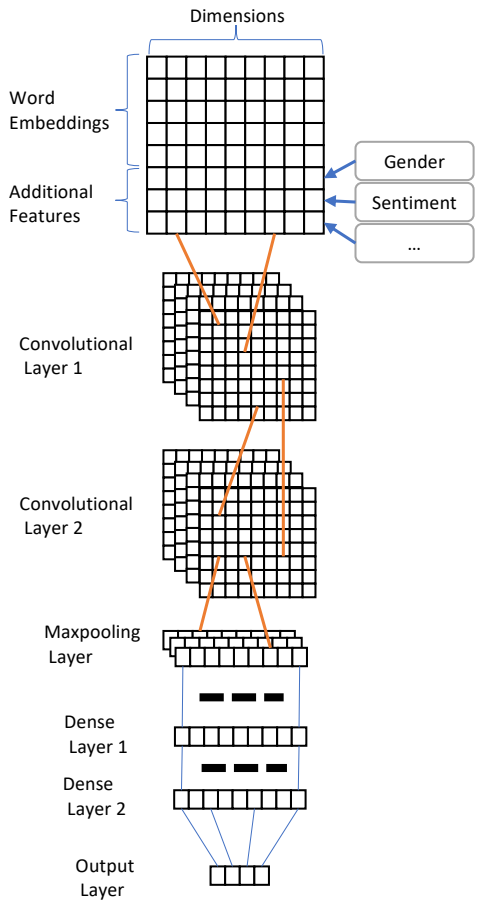

Figure 1: Architecture of the CNN implemented for the GermEval 2018 Shared Task.

closer to one another (see Figure 1).

As illustrated in the architectural overview, the additional features we experimented with are added to the training data in the pre-processing steps, the exact shape or form depending on the individual feature (i. e., binary values for gender or sentiment, embeddings for user descriptions, etc., see Section 4.1 for more details).

5 Results

The results presented below were obtained using cross-validation[5] on the training portion of the data set provided by the organisers of GermEval 2018. We compute the average accuracy for the binary classification (OFFENSE vs. OTHER) for Task 1 (Table 1) and provide accuracy, precision, recall and f1-score for the individual classes (INSULT, ABUSE, PROFANITY and OTHER) in Task 2 (Table 2). Based on the cross-validation over the training data, we consider the Twitter embeddings in combination with user descriptions to be the best setup, with an accuracy of 81 for Task 1 and 72.2 for Task 2. However, because this approach is dependent on the existence of user mentions in the tweet text, which may be proportionally less present in the test set, the figures on the test data may well deviate and show another setup to be the best performing one.[6]

6 Runs

We have submitted six runs (three for each task):

1. **dfkilt_coarse_1.txt:** TE+Desc approach including twitter embeddings and user mentions description (Task 1).

2. **dfkilt_coarse_2.txt:** TE+Sent approach including twitter embeddings and sentiment analysis information (Task 1).

3. **dfkilt_coarse_3.txt:** TE+G+D approach including twitter embeddings, gender classification and mentions descriptions (Task 1).

4. **dfkilt_fine_1.txt:** TE+Desc approach including twitter embeddings and mentions description (Task 2).

5. **dfkilt_fine_2.txt:** TE+S+G+D approach including twitter embeddings, sentiment analysis, gender classification and mentions description (Task 2).

6. **dfkilt_fine_3.txt:** TE+S+D approach including twitter embeddings, sentiment analysis and mentions description (Task 2).

7 Discussion

When dealing with the task of detecting hateful, aggressive, racist and/or sexist behaviour online, a lack of high inter-annotator agreement can be an issue and shows the high complexity of the challenge – even for humans. Ross et al. (2016) for instance introduce a German corpus of hate speech on the European refugee situation and report very low inter-annotator agreement scores (Krippendorff's α between 0.18 and 0.29). Waseem (2016) investigates inter-annotator agreement when comparing amateur annotations (generated using CrowdFlower)

[5]Due to time constraints we performed cross-validation with one single fold only.

[6]Note that we generated additional training data through user friends for the classes OFFENSE and OTHER only and, hence, did not use the data in Task 2.

Table 1: Results for Task 1 using different features

	Acc	OFFENSE			OTHER		
		P	R	F1	P	R	F1
Fasttext (FT)	73.9	–					
Wikipedia Embeddings (WE)	71.8	69.4	36.9	48.2	72.4	91.1	80.7
Twitter Embeddings (TE)	72.7	62.4	58.1	60.2	77.8	80.8	79.2
TE + Sentiment	78.5	**80**	52.5	63.4	78	**92.8**	84.8
TE + Descriptions	**81**	79	62.3	**69.6**	**81.8**	91.1	**86.2**
TE + Gender Classification	76.3	66.7	**66.3**	66.5	81.5	81.8	81.6
TE + Sentiment + Gender	75.4	66.2	62.5	64.3	80	82.5	81.2
TE + Sentiment + Descriptions	76.1	67.3	63.1	65.2	80.4	83.2	81.8
TE + Gender + Descriptions	76.9	72.2	56.9	63.6	78.8	88	83.1
TE + Sentiment + Gender + Descriptions	75.6	67.4	60.6	63.8	79.5	83.8	81.6
TE + User Friends Information	77.2	74.4	53.1	62	78.2	90.2	83.8

Table 2: Results for Task 2 using different features. (FT: Fasttext, WE: Wikipedia Embeddings, TE: Twitter embeddings, S: Sentiment, G: Gender Classification, D: Descriptions, UFI: User friends information)

	Acc	INSULT			ABUSE			PROFANITY			OTHER		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
FT	68.3	–											
WE	67.6	**39.5**	26.8	31.9	49.1	52	50.5	0	0	0	77.7	81.4	79.5
TE	65.2	27.3	5.4	9	41.3	57.8	48.2	0	0	0	78.1	79.7	79
TE+S	69	33.8	46.4	39.1	54.5	47.1	50.5	0	0	0	82.9	81.4	82.1
TE+D	**72.2**	38.9	**55.2**	**45.7**	62.7	42.4	50.6	0	0	0	**83.5**	86.8	**85.1**
TE+G	69.6	37.8	25	30.1	51.5	51	51.2	0	0	0	79.2	85.2	82.1
TE+S+G	70.1	28.6	28.6	28.6	65.7	45.1	53.5	0	0	0	78.2	87.3	82.5
TE+S+D	71.4	33.3	1.8	3.4	51.1	**64.7**	**57.1**	0	0	0	79.9	87.6	83.6
TE+G+D	69.2	30.9	44.6	36.5	63.5	32.4	42.9	0	0	0	79.9	87.3	83.4
TE+S+G+D	71.8	38.5	17.9	24.4	**70.6**	35.3	47.1	0	0	0	74.3	**95.5**	83.6
TE+UFI	—												

Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 21, 2018

and expert annotations and reports a similarly low Cohen's Kappa of 0.14. Van Hee et al. (2015) work on classification of cyberbullying using a Dutch corpus and report Kappa scores between 0.19 and 0.69. Kwok and Wang (2013) report an overall inter-annotator agreement of only 33% when investigating racist tweets. Nobata et al. (2016) report a relatively high agreement for binary classification of *clean* vs. *abusive* for social media comments on Yahoo! (Kappa = 0.843), but this number drops significantly when different subcategories for the abusive comments are introduced (such as *hate*, *derogatory language* and *profanity*, with Kappa decreasing to 0.456).

Using the basic setup of our network with Twitter embeddings does not improve over the FastText baseline (with accuracies of 72.2 vs. 73.9 for Task 1 and 65.2 vs. 68.3 for Task 2, respectively). However, adding additional types of information (or combinations), we do improve over this baseline, by 7.1 points in accuracy for Task 1 and 3.9 points in Task 2 in the best scoring setup.

In addition to different opinions on what constitutes and does not constitute "offensive language" (in terms of inter-annotator agreement), also the usage of automatically labelled data has its limitations. While 'distantly labelled' data might have a beneficial effect if manually labelled data is small, it might lose its effect with increasing gold standard data. The quality of automatically labelled data also plays an important role. As mentioned before, even Twitter users who post large numbers of offensive tweets do not do so exclusively. In various cases people might show a radical opinion without being explicitly offensive, and sometimes people also just talk about daily life using standard, acceptable language. Yet other times, they may use highly offensive language when complaining about the weather. The same rather high variance can be observed for people belonging to the OTHER user group. This means the data contains a large number of false positives and false negatives. A method which is able to deal with noisy data more robustly might have been more suitable.

Adding explicit sentiment information did improve over the setup using only the Twitter embeddings. Intuitively, a negative sentiment can be expected to align with the OFFENSE class for Task 1, and perhaps be less informative for Task 2. This is in any case reflected by the scores, as there is an almost 6 point increase in accuracy for Task 1, but a smaller increase for Task 2 (almost 4 points). However, a closer analysis shows, that many tweets might contain negative sentiment words without being offensive, such as 'arbeitslos' ('unemployed') or 'Flüchtling' ('refugee').

As for the added gender information, doing a factorized analysis of the different classes (in Task 1 and Task 2) and gender distribution, we did not see a clear hint that either male or female authors behave more offensive, profane, abusive or insulting. Yet, this feature improved performance for both tasks. While perhaps a clear correlation could not be established, it is possible that by including gender information, we are implicitly encoding certain features of tweets that help the network in differentiating between the classes.

Adding the descriptions that users publish about themselves (on their Twitter profile pages) increased the most when cross-validating the training set, compared to the setup using only embeddings. As explained in Section 4.1, the idea behind this feature is that certain users could be more likely to trigger hateful language. This would be captured by the classifier without the description as well (i. e., the user name showing up in the user mention would be an important feature). However, since user names are not likely to be present in the embeddings (hence, they will not have an informative representation using only the embeddings setup), adding the description the users added themselves, consisting of individual words which are more likely to be represented in the embeddings, will add information. For Task 1, this additional information improves just over 8 points to the embeddings-only setup, and for Task 2 the improvement is 7 points in accuracy.

Apart from the presented approaches, we made the first steps towards exploiting available resources in other languages to have at our disposal more training data for the neuronal networks. Given that the task is concerned with German tweets with limited amounts of German data available, we experimented with a crosslingual approach, i. e., expanding on the German language data by adding English language data. For the first attempt, we used the NLP+CSS_2017 data set (Jha et al., 2017).[7] The original data set (containing 10,095 unique tweets) was annotated for detecting benevolent sexism (labels used: BENEVOLENT,

[7] https://github.com/AkshitaJha/NLP_CSS_2017

HOSTILE, OTHER). Matching the definition of abusive language according to GermEval's annotation guidelines all instances of sexism found in the cleaned corpus (only tweets with a clear inner-annotator agreement were kept) were tagged as ABUSE and all remaining tweets were simply classified as OTHER.

In order to use data sets in different languages, we mapped the word embeddings of both data sets (one in English, another in German) onto each other, both generated from Wikipedia data, using MUSE.[8] Under the assumption that the specific characteristics (word embeddings) use the same vector space, the neural network should not explicitly register the difference between English and German training data, and should, hence, produce better results. This crosslingual approach produces 71.5% average accuracy in Task 1 and 67.9% average accuracy in Task 2. These preliminary results demonstrate that the accuracy numbers have not increased compared to the other approaches. We will investigate the crosslingual approach in more detail in follow-up work.

8 Conclusions and Future Work

We have developed a CNN-based approach on German Twitter data to predict offensiveness. The data is annotated at two levels; one coarse level indicating whether or not the tweet is offensive (Task 1), and one detailed level indicating whether offensive tweets are insulting, profane or abusive (Task 2). We augment the available training data with several different types of information and in the best scoring setup achieve an accuracy increase of 7.1 points for Task 1 and 3.9 points for Task 2, comparing to a baseline implementation using FastText. This sets the marks of our best attempt at an accuracy of 81 for Task 1 and 72.2 for Task 2.

Various previous studies and also our own experiments demonstrate that the automatic classification of offensive language, including closely related linguistic categories, with a very high degree of accuracy is a very challenging task. The low inter annotator agreement often mentioned above is, obviously, due to the highly subjective nature of language perception and interpretation. For some people certain expressions constitute "offensive language", for others they do not. It is challenging, maybe even impossible, to break this down into a binary classification task or into a task with a small number of categories. This socio-technical challenge notwithstanding, it is surely worthwhile to continue this line of research to arrive at larger data sets, better and more adequate categories and more suitable evaluation procedures. It would also be interesting to investigate the different ways an automatic text classification procedure could help and assist social media users flagging and responding to, but also composing messages. After all, maybe many instances of offensive language could be taken care of by making sure that they never come into existence. For example, Twitter users who are writing a tweet or a reply to a certain user and who use, based on an automatic classifier, offensive language, could be shown an alert window before posting, reminding them that they are probably using offensive language and that there is an actual human being on the other end of the line who may take offense by language of this nature.

[8] https://github.com/facebookresearch/MUSE

Acknowledgments

This work has been partially funded by the project LYNX. The project LYNX has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780602. More information is available online at http://www.lynx-project.eu.

References

Banks, James. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology, 24(3)*

Badjatiya, Pinkesh and Gupta, Shashank and Gupta, Manish and Varma, Vasudeva. 2017. Deep learning for hate speech detection in tweets *Proceedings of the 26th International Conference on World Wide Web Companion, 759–760*

Bourgonje, Peter and Moreno-Schneider, Julian and Srivastava, Ankit and Rehm, Georg 2017. Automatic classification of abusive language and personal attacks in various forms of online communication *International Conference of the German Society for Computational Linguistics and Language Technology, 180–191* Springer.

Burnap, Pete and Williams, Matthew L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 223–242* Wiley Online Library.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar. 2017. Automated hate speech detection and the problem of offensive language *arXiv preprint arXiv:1703.04009*

Harald Dreißing and Josef Bailer and Anne Anders and Henriette Wagner and Christine Gallas. 2014. Cyberstalking in a large sample of social network users: prevalence, characteristics, and impact upon victims. *Cyberpsychology, Behaviour, and Social Networking, 17(2)*

Djuric, Nemanja and Zhou, Jing and Morris, Robin and Grbovic, Mihajlo and Radosavljevic, Vladan and Bhamidipati, Narayan. 2015. Hate speech detection with comment embeddings *Proceedings of the 24th international conference on world wide web, 29–30*

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Van Hee, Cynthia and Lefever, Els and Verhoeven, Ben and Mennes, Julie and Desmet, Bart and De Pauw, Guy and Daelemans, Walter and Hoste, Veronique. 2015. Detection and Fine-Grained Classification of Cyberbullying Events. *Proceedings of the International Conference Recent Advances in Natural Language Processing, 672-680*

Jha, Akshita and Mamidi, Radhika 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. *Proceedings of the Second Workshop on NLP and Computational Social Science, 7-16*

Jones, Lisa M and Mitchell, Kimberly J and Finkelhor, David. 2013. Online harassment in context: Trends from three youth internet safety surveys (200, 2005, 2010). *Psychology of violence, 3(1):53* Educational Publishing Foundation.

Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Douze, Matthijs and Jgou, Hrve and Mikolov, Tomas. 2016. FastText.zip: Compressing text classification models.

Robin M. Kowalski and Susan P. Limber. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health, 53(1)*

Kwok, Irene and Wang, Yuzhou. 2013. Locate the Hate: Detecting Tweets Against Blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 1621-1622*

Alice E. Marwick and Ross W. Miller. 2014. Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape. *Fordham Center on Law and Information Policy Report*

Massaro, Toni M. 1991. Equality and Freedom of Expression: The Hate Speech Dilemma. *William & Mary Law Review, 32(211)*

Nand, Parma and Perera, Rivindu and Kasture, Abhijeet. 2016. "How Bullying is this Message?": A Psychometric Thermometer for Bullying *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 695–706*

Nobata, Chikashi and Tetreault, Joel and Thomas, Achint and Mehdad, Yashar and Chang, Yi. 2016. Abusive language detection in online user content *Proceedings of the 25th international conference on world wide web, 145–153*

Park, Ji Ho and Fung, Pascale. 2017. One-step and two-step classification for abusive language detection on twitter *arXiv preprint arXiv:1706.01206*

Phillips, Whitney. 2015. This Is Why We Can't Have Nice Things: Mapping the Relationship between Online trolling and Mainstream Culture. The MIT Press, Cambridge.

Pitsilis, Georgios K and Ramampiaro, Heri and Langseth, Helge. 2018. Detecting Offensive Language in Tweets Using Deep Learning *arXiv preprint arXiv:1801.04433*

Ross, Björn and Rist, Michael and Carbonell, Guillermo and Cabrera, Benjamin and Kurowsky, Nils and Wojatzki, Michael. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, 17:6-9*

Tatjana Scheffler 2014. A German Twitter Snapshot *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* European Language Resources Association (ELRA). Reykjavik, Iceland

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

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

Jin Wang and Zhongyuan Wang and Dawei Zhang and Jun Yan 2017. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2915–2921*

Warner, William and Hirschberg, Julia. 2012. Detecting hate speech on the world wide web *Proceedings of the Second Workshop on Language in Social Media, 19–26*

Waseem, Zeerak and Hovy, Dirk. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter *Proceedings of the NAACL student research workshop, 88-93*

Waseem, Zeerak. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science, 138-142* Educational Publishing Foundation.

Waseem, Zeerak and Davidson, Thomas and Warmsley, Dana and Weber, Ingmar. 2017. Understanding abuse: a typology of abusive language detection subtasks *arXiv preprint arXiv:1705.09899*

Ulli Waltinger 2010. GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)* electronic proceedings. Valletta, Malta

Wulczyn, Ellery and Thain, Nithum and Dixon, Lucas. 2017. Ex machina: Personal attacks seen at scale *Proceedings of the 26th International Conference on World Wide Web, 1391–1399*