

Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication

Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava,
and Georg Rehm^(✉)

Language Technology Lab, DFKI GmbH,
Alt-Moabit 91c, 10559 Berlin, Germany
georg.rehm@dfki.de

Abstract. The sheer ease with which abusive and hateful utterances can be made online – typically from the comfort of your home and the lack of any immediate negative repercussions – using today’s digital communication technologies (especially social media), is responsible for their significant increase and global ubiquity. Natural Language Processing technologies can help in addressing the negative effects of this development. In this contribution we evaluate a set of classification algorithms on two types of user-generated online content (tweets and Wikipedia Talk comments) in two languages (English and German). The different sets of data we work on were classified towards aspects such as racism, sexism, hatespeech, aggression and personal attacks. While acknowledging issues with inter-annotator agreement for classification tasks using these labels, the focus of this paper is on classifying the data according to the annotated characteristics using several text classification algorithms. For some classification tasks we are able to reach f-scores of up to 81.58.

1 Introduction

Hateful conduct, abusive language and verbal aggression are by no means new phenomena. Comments and statements of this type seriously hamper a constructive private discussion or public debate. The sheer ease with which hateful utterances can be made – typically from the comfort of your home and the lack of any immediate negative repercussions – using today’s digital communication technologies, is responsible for their significant increase and global ubiquity. In recent years, the topic has received an increasing amount of attention from multiple stakeholders. Among these are social scientists who want to analyse this phenomenon and reasons for abusive online behaviour and politicians who realise that major parts of public debates and social discourse are carried out online. In addition, we have seen that not only such online discussions but also the perception of concepts, politicians, elections and civil rights movements can be influenced using highly targeted social media marketing campaigns. We live in a time in which online media, including online news and online communication,

have an unprecedented level of social, political and also economic relevance. This situation creates a plethora of challenges with regard to the key question how best to address the importance and relevance of online media and online content with technological means while at the same time not putting in place a centralised infrastructure that can be misused for the purpose of censorship or surveillance. One challenge is to separate high quality content from offensive, hateful, abusive or massively biased content. While these tasks have been mostly in the realm of journalism, they are getting more and more transferred to the end user of online content, i.e., the analysis, curation and assessment of information is no longer carried out by professional news editors or journalists exclusively – the burden of fact checking is more and more left to the reader.

In the social sciences and humanities, research on the phenomena and characteristics of Computer-Mediated Communication (CMC) has a long tradition. Initially, scholars concentrated on different types of novel communication media such as electronic mail, Internet Relay Chat (IRC), Usenet newsgroups, and different types of hypertext systems and documents, especially personal home pages, guestbooks and, later, discussion fora (Runkehl et al. 1998; Crystal 2001; Storrer 2001; Döring 2002). Early on, researchers focused upon the obvious differences between these new forms of written digital communication and the established, traditional forms, especially when it comes to linguistic phenomena that can be observed on the text surface, such as smileys and emoticons, specific acronyms and technological aspects of communication. Many authors observed that the different forms of internet-mediated communication have a certain *oral* and *spoken* style, quality and conceptualisation to them, as if produced spontaneously in a casual conversation, while, at the same time, being realised in a *written* medium (Haase et al. 1997).

If we now fast forward to 2017, a completely different picture emerges. About 40–50% of the global population has access to the Internet, most of whom also use the World Wide Web and one or more of the big social networks. The internet has become mainstream and acts like an amplifier, maybe also as an enabler, of social trends. We already mentioned some of the current challenges of this massive penetration of our lives through Internet-based forms of communication. The social, political and economic relevance of online media, online news and online communication could not be any more crucial. While early analyses and discussions of computer-mediated communication and discourse, e.g., (Reid, 1991), observed that their participants were involved in the “deconstruction of boundaries” and the “construction of social communities”, today the exact opposite seems to be case: both offline and online can we observe the (disturbing) trend of increased nationalism and the exclusion of foreigners, immigrants and seemingly arbitrary minorities – boundaries are constructed, social communities deconstructed.

One last aspect is worth pointing out: up to now there has not really been any major need for automatic classification approaches of online content, with two notable exceptions. The first are online advertisements, either in the form of unsolicited spam email or in the form of online ads, either embedded in web

documents or presented as pop-out windows. The second exception is sentiment analysis of social media data, driven by a clear use case: knowing what your customers or voters say or think about you as a company or politician. We are now slowly approaching a state of play, in which automatic means may be needed to classify online content or parts of online content into additional dimensions such as, for example, “hatespeech”, “abusive language”, maybe even “fake news” and “alternative facts”. While spam mail can be classified and categorised with a fairly high accuracy (and online ads taken care of with an ad blocker), sentiment analysis already poses more difficult challenges (such as irony, sarcasm and scope issues). And it remains to be seen if abusive language can be identified accurately using automatic means and if additional analysis dimensions have to be incorporated if automatic approaches are to be used in a real application scenario.

The research presented in this paper has been carried out under the umbrella of a two-year research and technology transfer project. We collaborate with four SME partners that all face the challenge of having to process, to analyse and to make sense of large amounts of digital information. The four companies cover four different use cases and sectors (Rehm and Sasaki 2015), including journalism. For these partners we develop a platform that provides several semantic and knowledge technologies. In this article, we focus upon the application of several classification algorithms to establish the feasibility of the detection and classification of abusive language. We do this by evaluating the classification algorithms on three publicly available data sets. While the definition of abusive language, and, consequently, inter-annotator agreement of relevant corpora are crucial and far from resolved issues in this area of research, we focus on classification using linguistic features. Our goal is to establish a solid baseline for these three, publicly available corpora. The remainder of this paper is divided into four sections. Section 2 discusses related work, most notably on inter-annotator agreement, an important prerequisite for accurate classification. Section 3 describes the analysed and classified data sets. Section 4 discusses the results and Sect. 5 provides a summary and ideas for future work.

2 Related Work

Today, when it comes to the characteristics of online media and communication, several challenges are being discussed over and over again. We believe that the Natural Language Processing (NLP) and Language Technology (LT) communities can provide at least parts of the adequate social and technical solutions for, among others, hatespeech, fake news (including orchestrated disinformation campaigns), politically biased journalism, trolling, cyber-bullying and abusive language.

Hateful or aggressive conduct online has received an increasing amount of attention in recent years. For an accurate classification, however, consensus is needed on what constitutes abusive language, hate speech and aggressive conduct, and what is still acceptable within the boundaries of free speech.

While automated methods for detecting and classifying language use – such as spam vs. no spam – and sentiment – such as positive vs. negative – are typical application scenarios for NLP technologies, the question of what is acceptable and no longer acceptable within the boundaries of free speech puts this sensitive question and area of research into the intersection of different 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.

Many researchers approaching this topic acknowledge the difficulty in reaching a consensus. Ross et al. (2016) introduce a German corpus of hate speech on the European refugee crisis and report low inter-annotator agreement scores (Krippendorff’s α between 0.18 and 0.29). Waseem (2016) investigates inter-annotator agreement comparing amateur annotations using CrowdFlower and expert annotations using precise instructions and reports a 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) investigate racist tweets and report an overall inter-annotator agreement of only 33%. Nobata et al. (2016) report a relatively high agreement for binary classification of clean vs. abusive for Yahoo! comments (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).

Another complicating issue is the fact that abusive language is often extralinguistic in nature. Whether a particular utterance is considered abusive or not, often depends on other aspects including context, (ethnicity of the) author, (ethnicity of the) targeted person or group, etc. (Nand et al. 2016; Waseem and Hovy 2016; Warner and Hirschberg 2012). An excellent overview of NLP-based approaches towards hate speech detection is provided by Schmidt and Wiegand (2017).

In this paper we focus on the classification task and present several classification scores using multiple available data sets.

3 Data Sets

The experiments reported in this paper are conducted on three different data sets. The first one (ET, see Table 1) is provided by Waseem and Hovy (2016) and consists of English tweets. We scraped the tweets’ actual content; of the 16,907 tweet IDs provided on the authors’ GitHub page, we were able to retrieve 15,979 tweets (the smaller number most likely due to deleted tweets or time-outs during scraping). The tweets were classified into the classes *none* (10,939 instances), *sexism* (3,131 instances) or *racism* (1,909 instances). Despite missing out on 928 annotated tweets, the distribution over our version of the data set is the same as the one reported by Waseem and Hovy (2016), with respectively 68%, 20%, 12% of tweets being annotated as *none*, *sexist*, *racist*. For this and the other two data sets, Table 1 provides some examples.

Table 1. The three data sets – key facts and examples

Data Set: English Tweets (ET) – (Waseem and Hovy, 2016)	
	15,979 English language tweets (no. of words 273,805, avg.: 17)
Classes	<i>none</i> (10,939), <i>sexism</i> (3,131), <i>racism</i> (1,909)
Examples	@Fixer.Turkey Why were innocent civilians in prison? (<i>none</i>) @shaner38 Hope not. How will she pay her bills? (<i>sexism</i>) @FalconEye123456 May Allah bless him with 72 virgin pigs. (<i>racism</i>)
Data Set: German Tweets (GT) – (Ross et al., 2016)	
	469 German language tweets (no. of words: 8,650, avg.: 18)
Classes	<i>hateful</i> (104), <i>non-hateful</i> (365)
Examples	Deutsche Frauen an #rapefugees opfern. #wasistlinks (<i>hateful</i>) Flüchten, wo andere Urlaub machen. #Idomeni #refugeesnotwelcome (<i>non-hateful</i>) Gegen #Multikulti hab ich eigentlich nichts, gegen #Islamisierung schon. (<i>non-hateful</i>)
Data Set: Wikipedia Talk (WT) – (Wulczyn et al., 2016)	
	11,304 English language Wikipedia Talk comments (no. of words: 739,494, avg.: 65)
Classes	<i>aggression</i> (8,674) vs. <i>no aggression</i> (2,630) – <i>attack</i> (2,498) vs. <i>no attack</i> (8,806)
Examples	You stick to your talk page, I'll d mine, right? 20: (<i>none</i>) :::Yes, and Kudpung himself called for an admin's desysop in the section just above this one. What base hypocrisy. Perhaps he does not realize his own membership in his "anti-admin brigade", the existence of which he has never provided a shred of evidence for. (<i>attack</i>) == Thomas W == : Don't bother telling him anything. He'll cry to his butt buddy Bishonen, who happens to have admin powers. (<i>aggression</i>) == Suck it! == If you can't understand this common American idiom then perhaps you shouldn't be editing Wikipedia. At any rate, why are you monitoring my talk page, stalker? (<i>aggression</i>)

The second data set (GT, see Table 1) is provided by Ross et al. (2016) and consists of German tweets. With only 469 tweets, this data set is considerably smaller. They were annotated by two expert annotators who indicated a tweet to be either *hateful* or *not hateful*. In addition, the second annotator also scored the tweet on a scale of 1 (not offensive at all) to 6 (very offensive). The distribution of *hateful* vs. *non-hateful* tweets for annotator 1 was 110–359 and for annotator 2 it was 98–371.

The third data set (WT, see Table 1) is described by Wulczyn et al. (2016) and consists of user comments on Wikipedia Talk pages. This corpus is annotated for toxicity, aggression and personal attacks on users; the annotations are obtained

through crowd-sourcing. Due to the sheer size of the complete data set, we only downloaded part of it (user comments from 2013 to 2015)¹ as well as the annotations for aggression and personal attacks.² This resulted in 11,304 annotated comments, 8,806 were annotated as cases of an *attack* and 2,498 as cases of *no attack*. 2,630 comments were annotated as containing *aggression* and 8,676 as *no aggression*. In the case of aggression, a rating was annotated as well. On a scale of -3 (very aggressive) to 3 (very friendly),³ the distribution from -3 to 3 was as follows: -3 : 772; -2 : 635; -1 : 1,223; 0 : 7,623; 1 : 717; 2 : 243; 3 : 91.

4 Evaluation

We applied a set of classification algorithms (Bayes, Bayes expectation maximization, C4.5 Decision Trees, Multivariate Logistic Regression, Maximum Entropy and Winnow2) on all three corpora using the Mallet Machine Learning for Language toolkit (McCallum 2002).⁴ All classifiers use a Bag of Words (BOW) feature set (word unigrams). The figures in Table 2 are the result of

Table 2. Results of our classification experiments

	Bayes	Bayes exp. max.	C4.5	Logistic Regression	Maximum Entropy	Winnow2		Bayes	Bayes exp. max.	C4.5	Logistic Regression	Maximum Entropy	Winnow2
	English Tweets (ET)							German Tweets (GT) – (binary, exp. 1)					
<i>accuracy</i>	84.61	84.01	82.95	85.67	83.67	76.66	<i>accuracy</i>	75.74	78.93	74.04	77.23	75.96	71.91
<i>precision</i>	80.54	79.57	79.07	83.57	81.20	69.85	<i>precision</i>	70.65	75.07	69.30	74.80	72.46	72.41
<i>recall</i>	78.63	77.97	74.37	77.45	74.37	69.62	<i>recall</i>	74.78	76.06	74.98	76.58	74.85	72.68
<i>f-score</i>	79.10	78.34	76.17	80.06	77.20	69.32	<i>f-score</i>	65.84	69.74	70.66	71.98	73.02	71.15
	German Tweets (GT) – (binary, exp. 2)							German Tweets (GT) – (rating)					
<i>accuracy</i>	80.21	74.26	76.81	79.15	76.38	77.23	<i>accuracy</i>	36.60	35.32	37.87	33.40	34.89	25.53
<i>precision</i>	72.76	73.59	72.54	77.18	73.62	74.65	<i>precision</i>	42.51	39.76	56.22	31.39	31.90	38.17
<i>recall</i>	77.57	79.49	77.85	79.74	77.31	76.37	<i>recall</i>	38.53	38.19	38.76	36.34	35.71	25.84
<i>f-score</i>	70.93	68.97	69.85	75.41	74.20	73.05	<i>f-score</i>	27.43	27.03	23.68	30.34	30.75	24.06
	Wikipedia (WT) – Attack (binary)							Wikipedia (WT) – Aggression (binary)					
<i>accuracy</i>	83.11	82.70	81.08*	80.90	77.71	77.77	<i>accuracy</i>	82.19	82.10	79.58*	80.42	77.17	79.08
<i>precision</i>	81.78	81.33	79.27*	79.36	76.03	77.11	<i>precision</i>	80.68	80.60	78.13*	78.91	75.26	77.25
<i>recall</i>	83.14	82.83	81.31*	80.97	77.87	77.83	<i>recall</i>	82.01	81.87	80.18*	80.46	77.29	78.57
<i>f-score</i>	81.58	81.36	79.27*	79.74	76.65	77.28	<i>f-score</i>	80.60	80.57	78.37*	79.23	75.80	77.45
	Wikipedia (WT) – Aggression (rating)												
<i>accuracy</i>	67.13	67.40	66.81*	65.28	57.77	55.73							
<i>precision</i>	57.21	56.05	54.08*	57.42	57.21	54.07							
<i>recall</i>	67.27	66.94	66.42*	65.68	58.18	55.73							
<i>f-score</i>	59.13	59.00	58.14*	59.95	55.26	54.53							

¹ https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973.

² https://figshare.com/projects/Wikipedia_Talk/16731.

³ While the documentation states a range from -2 to 2 , we actually found a range of -3 to 3 in the annotations (https://meta.wikimedia.org/wiki/Research:Detox/Data_Release).

⁴ <http://mallet.cs.umass.edu/api/>.

ten-fold cross-validation⁵ with a 90–10 distribution of training and test data. Note that the table incidentally contains f-scores that are lower than their corresponding precision and recall scores due to averaging over the precision, recall and f-scores for every class that exists in the data.

While Waseem and Hovy (2016) report better results for character n -grams compared to word n -grams (73.89 vs. 64.58) on their data set (ET), Mallet’s logistic regression implementation, using word unigrams, outperforms the best scoring feature set in Waseem and Hovy (2016) (i. e., 80.06 vs. 73.89). The influence of using character n -grams vs. word n -grams may be language dependent. Nobata et al. (2016) report better performance with character n -grams, while Van Hee et al. (2015) report better performance with word n -grams (on a Dutch corpus) but in the above example, the same language and data is used. The type of classifier may also influence the features that are used, but Waseem and Hovy (2016) also use a logistic regression classifier. We have experimented with other features, such as word-ngrams, character-ngrams, and for the tweets cleaned the content using a set of regular expressions,⁶ but the best results were obtained with BOW features, as reported in Table 2. In addition, the most informative features are shown in Fig. 1.

The data set of Ross et al. (2016) is significantly smaller (GT). The annotations were done by two expert annotators. We have trained the set of classifiers twice, using these annotations (binary, expert 1 and binary, expert 2), and have made no attempts to resolve a final or definitive label, exactly because of the problems with inter-annotator agreement. While the results for the best scoring algorithm in the case of binary classification is still reasonable, performance

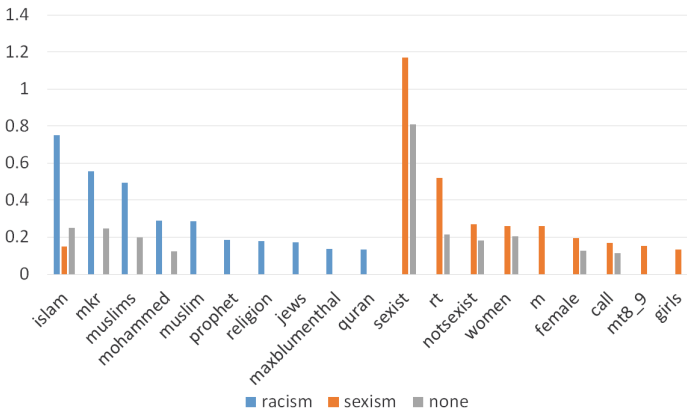


Fig. 1. The ten most informative features for classes in the (Waseem and Hovy, 2016) data set

⁵ Except for the numbers marked with an asterisk; these are the result of three-fold cross-validation due to the large amount of time needed for training and execution.

⁶ The set we used is inspired on the script available at <https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb>.

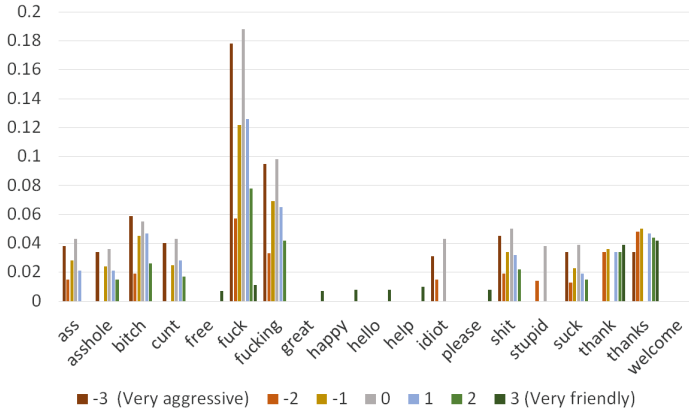


Fig. 2. The ten most informative features for classes in the Wikipedia talk aggression ratings data set

drops dramatically when using multi-label classification (six classes). The figures for the GT data set are of limited reliability due to its small size. Ross et al. (2016) do not report classification results and focus on inter-annotator agreement. We include the results in Table 2, but due to large score deviations for individual runs and for reasons of brevity, do not discuss the results for this data set.

For our subsection of the WT data set (Wulczyn et al., 2016), we see a similar pattern in the sense that binary classification scores are reasonable (81.58 and 80.60 for Bayes) but performance drops significantly when using multi-class classification (seven classes).

Wulczyn et al. (2016) do not mention any f-scores, but they do refer to experiments using their best performing personal attack classifier on comments made in 2015 (excluding “administrative comments and comments generated by bots”). The different setup in this study makes it hard to compare, as they are not performing binary classification directly, but assign a continuous score between 0 and 1. As the authors note, “even though the thresholded model-scores give good estimates of the rate of attacks over a random sample of comments, it is not given that they also give accurate estimates when partitioning comments into different groups”. Using this method, however, the authors do report a precision of 0.63 and a recall of 0.63, when using a threshold value of 0.425.

In an attempt to get a better overview of which classification algorithms work best for which type of data, we found that the preferred classification algorithm is highly data-specific. For example, Ranawana and Palade (2006) provide an overview of multi-class classification algorithms and identify the most distinct features in order to combine several algorithms for one classification problem. In their experiments, they work on a data set of human DNA from the Berkeley Drosophila Genome Project website. Andreopoulos et al. (2009) describes a variety of clustering algorithms applied to the biomedical domain. Caruana and

Niculescu-Mizil (2006) present an empirical study of several learning algorithms and their features applied on 11 data sets. It remains difficult to predict the best-performing algorithm when certain data characteristics (like size of the data set, average text length, variation, number of classes, etc.) are known. The main goal of the work presented in this paper is to provide a baseline performance indication and give an overview of different classification algorithms applied on the data sets included. We only use a simple BOW approach and the job of feature engineering is left as an important next step towards classification of abusive language.

To gain more insight into the features used by the best scoring algorithm for the different data sets, we extracted the ten most informative features and report scores averaged over ten runs. The x -axis contains the top 10 unique words for all existing classes. The y -axis represents the information gain per feature. Features for the first data set are shown in Fig. 1 for the three existing classes. The only overlap between the “racism” and “sexism” features (in the top 10) is *islam*, which was apparently also a relatively frequent term in the tweets classified as “sexist” in the training set. The considerable overlap between the “none” class and the other two is likely to reflect the fact that not all tweets containing these words were annotated as either “racist” or “sexist”. As also noted by Warner and Hirschberg (2012), classification of abusive language can be conceptualised as a word disambiguation task: sometimes the usage of a word is abusive, sometimes it is not. The features that are less straightforward as words represent the name of a cooking show (*mk*),⁷ the established abbreviation of “retweet” (*rt*), the twitter handle of a user (*mt8.9*)⁸ and (probably) the result of Mallet tokenising the m in $I'm$ as a separate word.

The ten most informative features for the classification task of *aggression* on a scale of -3 (very aggressive) to 3 (very friendly) in the WT data set for the best performing algorithm (Bayes) are shown in Fig. 2.⁹ Remarkably, the top 10 most informative features (words) for the seven classes are represented by only 18 unique words. The words in this set associated with friendliness (“welcome”, “please”, “great” and “happy”, for example) are only present in class 3. On the other end of the spectrum, class -3 only contains one word often associated with friendliness (“thanks”). Overall, there is a large degree of overlap between the classes, suggesting that the lower-ranked features also play an important role in classification. Upon manual investigation of the lower-ranked features, we found that the neutral class (0) seemed to function mostly as background noise, canceling out both extremes, as is the case for the classifier for the Waseem and Hovy (2016) data set. The negative digit classes (-1 , -2 and -3) contain a large vocabulary of curse words, whereas the positive digit classes (1, 2 and 3) contain more communicative, constructive and cooperative terms like “ask”, “questions”, “discussion”, etc.

⁷ https://en.wikipedia.org/wiki/My_Kitchen_Rules.

⁸ <https://twitter.com/mt8.9>.

⁹ Due to the small corpus size of the GT data set, we refrain from showing the ten most informative features here.

5 Conclusion

We apply a range of classification algorithms on three data sets that differ in language (usage), size and domain/genre: A corpus of English tweets annotated for racist and sexist tweets (ET); a corpus of German tweets annotated for hate speech (GT); and a corpus of English Wikipedia user comments who, unlike tweets, have less strict length restrictions (WT). While many studies on this topic focus on inter-annotator agreement (Sect. 2), we establish a baseline for classification based on these three corpora and data sets. We describe the differences between the performance of different classification algorithms and the features used by the respective best performing algorithm. Although consensus on what needs to be detected, which is typically reflected by inter-annotator agreement, is important to construct relevant training corpora, our results indicate that automatic classification can provide reasonable results and does not have to be a bottle neck in attempts to automatically detect abusive language. Nevertheless, further research is needed to reach an agreement on definitions of abusive language, hate speech, hateful conduct, cyber-bullying and other phenomena of online communication in order to arrive at accurate and meaningful classification approaches. Additionally, the simple feature set (BOW) used in the experiments can and should be complemented with more semantically and context-aware components such as sentiment values, dependency parsing (to provide insight on scope of abusive elements), and other more sophisticated NLP techniques. Finally, we would like to emphasise that the extra-linguistic nature of abusive behaviour and the fact that, apart from language usage, accurate detection depends on the author, target audience, communicative intent and other context specifics, is not addressed in this paper and, thus, left as an important part of future work in this area.

Acknowledgments. The authors wish to thank the anonymous reviewers for their helpful feedback. The project “Digitale Kuratierungstechnologien (DKT)” is supported by the German Federal Ministry of Education and Research (BMBF), “Unternehmen Region”, instrument “Wachstums Kern-Potenzial” (No. 03WKP45). More information at <http://www.digitale-kuratierung.de>.

References

- Andreopoulos, B., An, A., Wang, X., Schroeder, M.: A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics* **10**(3), 297–314 (2009). <https://doi.org/10.1093/bib/bbn058>. ISSN 1477-4054
- Banks, J.: Regulating hate speech online. *Int. Rev. Law Comput. Technol.* **24**(3), 233–239 (2010)
- Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161–168. ACM (2006)
- Crystal, D.: *Language and the Internet*. Cambridge University Press, Cambridge (2001)
- Döring, N.: Personal home pages on the web: a review of research. *J. Comput.-Mediat. Commun.* **7**(3) (2002). <http://www.ascusc.org/jcmc/>

- Dreifling, H., Bailer, J., Anders, A., Wagner, H., Gallas, C.: Cyberstalking in a large sample of social network users: prevalence, characteristics, and impact upon victims. *Cyberpsychology Behav. Soc. Netw.* **17**(2), 61–67 (2014)
- Haase, M., Huber, M., Krumeich, A., Rehm, G.: Internetkommunikation und Sprachwandel. In: Weingarten, R. (ed.) *Sprachwandel durch Computer*, pp. 51–85. Westdeutscher Verlag, Opladen (1997). https://doi.org/10.1007/978-3-322-91416-3_3
- Jones, L.M., Mitchell, K.J., Finkelhor, D.: Online harassment in context: trends from three youth internet safety surveys (2000, 2005, 2010). *Psychol. Violence* **3**(1), 53 (2013)
- Kowalski, R.M., Limber, S.P.: Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *J. Adolesc. Health* **53**(1), S13–S20 (2013)
- Kwok, I., Wang, Y.: Locate the hate: detecting tweets against blacks. In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2013*, pp. 1621–1622. AAAI Press (2013). <http://dl.acm.org/citation.cfm?id=2891460.2891697>
- Marwick, A.E., Miller, R.W.: *Online harassment, defamation, and hateful speech: a primer of the legal landscape*. Fordham Center on Law and Information Policy Report, June 2014
- Massaro, T.M.: Equality and freedom of expression: the hate speech dilemma. *William Mary Law Rev.* **32**(211) (1991)
- McCallum, A.K.: MALLETT: a machine learning for language toolkit (2002). <http://www.cs.umass.edu/~mccallum/mallet>
- Nand, P., Perera, R., Kasture, A.: “How bullying is this message?”: a psychometric thermometer for bullying. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 695–706. The COLING 2016 Organizing Committee, December 2016. <http://aclweb.org/anthology/C16-1067>
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Republic and Canton of Geneva, Switzerland*, pp. 145–153. International World Wide Web Conferences Steering Committee (2016). <https://doi.org/10.1145/2872427.2883062>, ISBN 978-1-4503-4143-1
- Phillips, W.: *This Is Why We Can’t Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. The MIT Press, Cambridge (2015)
- Ranawana, R., Palade, V.: Multi-classifier systems: review and a roadmap for developers. *Int. J. Hybrid Intell. Syst.* **3**(1), 35–61 (2006). <http://dl.acm.org/citation.cfm?id=1232855.1232859>, ISSN 1448-5869
- Rehm, G., Sasaki, F.: Digitale Kuratierungstechnologien - Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte. In: *Proceedings of the 2015 International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015*, pp. 138–139 (2015)
- Reid, E.M.: *Electropolis: communication and community on internet relay chat*, Honours thesis, University of Melbourne, Department of History (1991). <http://www.aluluei.com/electropolis.htm>
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: *Beißwenger, M., Wojatzki, M., Zesch, T. (eds.) Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochumer Linguistische Arbeitsberichte, Bochum*, vol. 17, pp. 6–9, September 2016
- Runkehl, J., Schlobinski, P., Siever, T.: *Sprache und Kommunikation im Internet - Überblick und Analysen*. Westdeutscher Verlag, Opladen (1998)

- Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, pp. 1–10. Association for Computational Linguistics, April 2017. <http://www.aclweb.org/anthology/W17-1101>
- Storrer, A.: Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation. In: Lehr, A., Kammerer, M., Konerding, K.-P., Storrer, A., Thimm, C., Wolski, W. (eds.) Sprache im Alltag. Beiträge zu neuen Perspektiven der Linguistik, pp. 439–466. de Gruyter, Berlin (2001). Herbert Ernst Wiegand zum 65. Geburtstag gewidmet
- Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., Hoste, V.: Detection and fine-grained classification of cyberbullying events. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 672–680. INCOMA Ltd., Shoumen (2015). <http://aclweb.org/anthology/R15-1086>
- Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, LSM 2012, Stroudsburg, PA, USA, pp. 19–26. Association for Computational Linguistics (2012). <http://dl.acm.org/citation.cfm?id=2390374.2390377>
- Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas, pp. 138–142. Association for Computational Linguistics, November 2016. <http://aclweb.org/anthology/W16-5618>
- Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, San Diego, California, pp. 88–93. Association for Computational Linguistics, June 2016. <http://www.aclweb.org/anthology/N16-2013>
- Wulczyn, E., Thain, N., Dixon, L.: Ex machina: personal attacks seen at scale. CoRR, abs/1610.08914 (2016). <http://arxiv.org/abs/1610.08914>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

