Detection of Abusive Language: the Problem of Biased Datasets

Michael Wiegand[*0], **Josef Ruppenhofer**[†‡], **Thomas Kleinbauer**[*]
[*]Spoken Language Systems, Saarland University, Saarbrücken, Germany
[†]Leibniz ScienceCampus, Heidelberg/Mannheim, Germany
[‡]Institute for German Language, Mannheim, Germany
michael.wiegand@lsv.uni-saarland.de
ruppenhofer@ids-mannheim.de
thomas.kleinbauer@lsv.uni-saarland.de

Abstract

We discuss the impact of data bias on abusive language detection. We show that classification scores on popular datasets reported in previous work are much lower under realistic settings in which this bias is reduced. Such biases are most notably observed on datasets that are created by focused sampling instead of random sampling. Datasets with a higher proportion of implicit abuse are more affected than datasets with a lower proportion.

1 Introduction

Abusive or offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Examples are (1)-(3). In the literature, closely related terms include *hate speech* (Waseem and Hovy, 2016) or *cyber bullying* (Zhong et al., 2016). While there may be nuanced differences in meaning, they are all compatible with the general definition above.

(1) stop editing this, you **dumbass**.
(2) Just want to slap the **stupid** out of these **bimbos**!!!
(3) Go lick a pig you arab muslim piece of **scum**.

Due to the rise of user-generated web content, in particular on social media networks, the amount of abusive language is also steadily growing. NLP methods are required to focus human review efforts towards the most relevant microposts.

In this paper, we examine the issue of data bias. For the creation of manually annotated datasets, randomly sampling microposts from large social media platforms typically results in a too small proportion of abusive comments (Wulczyn et al., 2017; Founta et al., 2018). Therefore, more focused sampling strategies have to be applied which cause biases in the resulting datasets. We show what implications this has on classifiers trained on these datasets: Previous evaluations reported high classification performance on datasets with difficult cases of abusive language, e.g. implicit abuse (§2). Contrarily, we find that the high classification scores are likely to be the result of modeling the bias in those datasets.

Although we will explicitly name shortcomings of existing individual datasets, our paper is not intended as a reproach of those who created them. On the contrary, we acknowledge the great efforts the researchers have taken to provide these resources. Without them, much existing research would not have been possible. However, we also noticed a lack of awareness of the special properties of those datasets among researchers using them. As we will illustrate with specific examples, this may result in unforeseen results of particular classification approaches.

2 Explicit and Implicit Abuse

One major distinction that has been proposed in the literature is the division into **explicitly** and **implicitly** abusive language (Waseem et al., 2017). The former are microposts that employ some abusive words (1)-(3) (e.g. *dumbass* or *scum*), while the latter represents the more difficult case in which the abusive nature is conveyed by other means, such as sarcasm, jokes, and particularly the usage of negative stereotypes etc. (4)-(5).

(4) i havent had an intelligent conversation with a woman.
(5) Jews don't marry children. Muslims do. All the time.

To determine which of the datasets that we consider in this work contain which type of abusive language, we proceeded as follows. On the set of abusive microposts of each dataset, we computed the proportion of microposts that include at least

[0]Present affiliation: Leibniz ScienceCampus, Heidelberg/Mannheim, Germany
[1]http://thelawdictionary.org/

Proceedings of NAACL-HLT 2019, pages 602–608
Minneapolis, Minnesota, June 2 - June 7, 2019.

one abusive word according to the lexicon of abusive words from Wiegand et al. (2018a). Datasets with a high proportion of abusive words typically contain a high amount of explicitly abusive microposts, whereas datasets with a low proportion contain a higher amount of implicitly abusive language. The resulting figures, of course, are only a lower bound estimate for explicit language abuse. There will also be microposts containing abusive words that are missing from the lexicon. However, after manual inspection of a sample of microposts, we are fairly confident that this does not significantly change the *relative* order of datasets when ranked according to their degree of explicit language abuse.

3 Datasets and Their Properties

Due to the limited space of this paper, we restrict our discussion to frequently cited (publicly available) datasets and datasets from shared tasks. Substantial interannotation agreement has also been reported with these datasets.

As we focus on the detection of abusive language in general, for those datasets containing more fine-grained class inventories describing subtypes of abusive language[2], we conflate the categories to one general category. As a result, there are always only two categories: *abuse* and *no-abuse*. This merging removes differences between the individual annotation schemes that would otherwise impede a meaningful comparison.

Table 1 shows a brief summary of the different datasets. Among the properties, we list the performance of a text classifier in the right-most column. Since in previous work performance on the different datasets was reported on the basis of different types of classifiers and also varying evaluation metrics, we ran the same classifier on all datasets in order to ensure a meaningful comparison. We chose FastText, which is an efficient supervised classifier known to produce state-of-the-art performance on many text classification tasks[3] (Joulin et al., 2017) and whose results are easy to reproduce. Performance is evaluated in a 10-fold crossvalidation setting using the macro-average F1-score.

Table 1 also describes the way the datasets were sampled. Not a single dataset has been produced by *pure* random sampling. This would always result in tiny proportions of abusive language. For example, Founta et al. (2018) estimate that on Twitter, there are only between 0.1% up to at most 3% abusive tweets. What comes closest to random sampling is the procedure followed by Founta et al. (2018), Razavi et al. (2010) and the Kaggle-challenge.[4] They took a random sample and applied some heuristics in order to boost the proportion of abusive microposts. For instance, in the Kaggle-challenge, further microposts from users were added who had been blocked due to being reported to post personal attacks.

The procedures applied by other researchers are more drastic because, as we show in §4 and §5, they affect more heavily the topic distribution of the dataset. These approaches do not even start with a random sample. The topic distribution is mostly determined by the creators of the dataset themselves. For example, Waseem and Hovy (2016) extract tweets matching query words likely to co-occur with abusive content. Kumar et al. (2018) choose Facebook-pages covering topics that similarly coincide with abusive language. The resulting datasets are far from representing a natural sample of the underlying social-media sites.

Table 1 shows that datasets that apply biased sampling (*Warner*, *Waseem*, *Kumar*) contain a high degree of implicit abuse. Boosted random sampling, which provides a more realistic cross section of microposts, on the other hand, captures a larger amount of explicit abuse. Future work should explore whether this is due to the predominance of explicit abuse on social media or some other reason, for example, the fact that human annotators more readily detect explicit abuse.

Intuitively, one would expect that the lower the proportion of explicit abuse is on the set of abusive microposts of a dataset, the lower the F1-score becomes because implicit abuse is not conveyed by lexical cues that are easy to learn. Table 1 confirms this notion, yet *Waseem* is the notable exception. We need to find an explanation for this deviation since *Waseem* is by far the most frequently used dataset for detecting abusive language (Badjatiya et al., 2017; Bourgonje et al., 2017; Pitsilis et al., 2018; Agrawal and Awekar, 2018; Karan and Snajder, 2018; Kshirsagar et al.,

[2]For example, Waseem and Hovy (2016) distinguish between *sexism* and *racism*.

[3]More involved classifiers achieve better performance, however, the relative differences between the datasets remain.

[4]www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

name	publication	source	microposts	%abusive	sampling	%explicit*	F1
Kaggle†	(Wulczyn et al., 2017)	Wikipedia	312,737	9.6	boosted random sampling	76.9	88.2
Founta	(Founta et al., 2018)	Twitter	59,357	14.1	boosted random sampling	75.9	87.3
Razavi	(Razavi et al., 2010)	diverse	1,525	31.9	boosted random sampling	64.7	83.3
Warner	(Warner and Hirschberg, 2012)	diverse	3,438	14.3	biased sampling	51.3	71.8
Waseem	(Waseem and Hovy, 2016)	Twitter	16,165	35.3	biased sampling	44.4	**80.5**
Kumar	(Kumar et al., 2018)	Facebook	15,000	58.1	biased sampling	32.7	70.4

Table 1: Properties of the different datasets. (*: *proportion of explicitly abusive microposts among abusive microposts.* †: *This is an extension of the dataset presented in Wulczyn et al. (2017). Details on the corpus creation about* Kaggle *can therefore be found in that publication.*)

2018; Mishra et al., 2018a,b; Park et al., 2018; Qian et al., 2018; Sahlgren et al., 2018; Sharifirad et al., 2018; Unsvåg and Gambäck, 2018; Wiegand et al., 2018a). This investigation is only possible since, fortunately, *Waseem* is one of the datasets whose creation process has been meticulously documented.

4 Topic Bias

The *Waseem*-dataset has been sampled in such a way that it contains a high proportion of microposts discussing the role of women in sports, particularly their suitability as football commentators. Such microposts also very often coincide with sexist remarks. However, the authors did not make any attempt to *debias* their dataset. As a consequence, domain-specific expressions such as *announcer*, *commentator*, *football* or *sport* occur very frequently and almost exclusively in abusive tweets. Yet intuitively these words should not be representative of abusive language. There are many texts on the web including Twitter that contain mentions of these expressions but that are not abusive. The current dataset, however, does not reflect that.

Table 2 illustrates this bias by listing the words with the highest Pointwise Mutual Information (PMI) towards abusive microposts. It compares the *Founta*-dataset (a dataset representing almost random sampling) with the *Waseem*-dataset (a dataset produced by biased sampling). We deliberately chose two datasets sampled from the *same* social-media site, namely Twitter, as otherwise the difference we report could be ascribed to differences in the underlying text sources. Table 2 shows that on the *Founta*-dataset, abusive words occupy the high ranks. Most of the highly ranked words of the *Waseem*-dataset, however, are not abusive. Similar observations can be made on the other datasets produced by biased sampling (i.e. *Warner* and *Kumar*). In the *Warner*-dataset, the

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

Table 2: Top 10 words having strongest correlation with abusive microposts according to PMI on *Founta* (dataset representing almost random sample) and *Waseem* (dataset produced by biased sampling).

Feature Set	Prec	Rec	F1
all words	80.91	80.08	80.49
(ii) query words removed	76.65	76.02	76.33
(i) topic words removed	75.07	74.41	74.72

Table 3: Impact of removing specific words from classifier trained and tested on *Waseem*.

words *CBS* and *Hollywood* are two of the most predictive words. They refer to the anti-semitic prejudice that Jews are supposed to control most of the US media. On that dataset, the bias of identity terms is also extreme: Almost 80% of the 256 mentions of the identity term *Jew* occur in abusive microposts. On the *Kumar*-dataset, even common Arabic person names, such as *Azan* or *Nahid*, strongly correlate with abusive language.

In order to demonstrate the detrimental effects such biases have, we now report the performance of further classifiers trained on the *Wassem*-dataset. Similar results could be obtained on the *Warner*- and *Kumar*-dataset. Yet they are most pronounced on the *Waseem*-dataset, which is also the dataset on which unexpectedly high classification performance has been observed in Table 1. Presumably, it is also the most biased dataset.

In our first experiment, we tested a FastText-classifier (§3) trained on the *Waseem*-dataset on a random sample of 500 additional tweets that include mentions of the topic words *football* and

racist		sexist	
author name	**freq**	**author name**	**freq**
Vile_Islam	1915	Yes You're Sexist	1320
Yes You're Sexist	8	Male Tears #4648	948
Standing Up 4 Trump	5	Vile_Islam	50
YESMarriageEquality	1	LilBeasy91	10
LilBeasy91	1	N!ck	9

Table 4: The 5 most sexist and racist authors on the *Waseem*-dataset and the number of their microposts.

sport. One would expect a low proportion of these particular tweets to be predicted as abusive. However, due to the fact that the abusive training data have such a large topic bias towards sports, the proportion of tweets predicted to be abusive is unreasonably high (i.e. 70%). Manual inspection confirmed that only a small proportion (up to 5%) was actually abusive. This result shows us that classifiers trained on the *Waseem*-dataset hardly generalize to the concept of abusive language. Difficult tweets on that dataset, e.g. instances of implicit abuse, may be classified correctly just because biased words such as *football* or *sport* occur in them.

In our second experiment, we train and test a classifier on the original *Waseem*-dataset in 10-fold crossvalidation. However, we remove either of the two types of biased words from the dataset:

(i) We remove 25 topic words from the 100 most correlating words that we thought bear no relation towards abusive language (e.g. *announcer*, *commentator*, *football* or *sport*).

(ii) We remove the 17 words that were used as a query by Waseem and Hovy (2016) to produce the dataset.

With (i) we want to show how good classifiers are that do not have access to biased words. This would be a realistic setting since words, such as *football* or *sport*, only have this bias towards abusive language on the *Waseem*-dataset. Such removal is also necessary since otherwise these biased words cause a huge amount of false positives when testing on other datasets (as shown above).

With (ii) we want to show that query words themselves are biased, too. For example, we observed that the query word *WomenAgainstFeminism* correlates with abusive tweets while *gamergate* correlates with non-abusive tweets. The purpose of query words is to retrieve tweets that address specific topics. The fact that they correlate with the classes of the dataset further proves that the focused sampling process introduces data bias.

The results of these two configurations are displayed in Table 3. It shows that the removal of a very few words (i.e. 0.2% of the overall vocabulary) already causes the classification performance to drop notably. Please note that these experiments do not capture the full impact of the bias in this dataset. That is, there will be more biased words beyond the 25 words we identified on the list of top 100 words ranked according to PMI since the cut-off value of 100 was arbitrarily chosen.

5 Author Bias

Datasets may also be affected by author bias. By that, we mean that information relating to the author of a micropost may artificially boost classification performance. Author information can be explicitly derived from *meta-information* of a micropost, for example, a feature that encodes the user name of a particular tweet that is to be classified. However, even if we do not explicitly encode such information, a (lexical) supervised classifier, such as the FastText-classifier from Table 1, may indirectly be affected by author biases. If the set of tweets belonging to a certain class predominantly comes from the same author, then a supervised classifier may largely pick up the writing style or the topics addressed by that author. Whenever the writing style or those topics are recognized, abusive language is predicted. This may work on a biased dataset but not beyond it.

We found that the distribution of abusive tweets on the *Waseem*-dataset is highly skewed towards 3 different authors as shown in Table 4. More than 70% of the sexist tweets originate from the two authors *Male Tears #4648* and *Yes, They're Sexist*. 99% of the racist tweets originate from a single author (i.e. *Vile_Islam*). If virtually all racist tweets originate from the same author, a classifier just needs to consider tweets from that author and can predict tweets from every other author as non-racist. On this particular dataset, such a strategy leads to good results: Both Qian et al. (2018) and Mishra et al. (2018a) proposed classification approaches that add author information to common text-level features. These approaches were solely evaluated on the *Waseem*-dataset. However, the author distribution on the *Waseem*-dataset does not reflect reality where abusive tweets originate from far more than a very few authors. In reality, we therefore should expect author information to be less predictive.

6 How to Avoid a Biased Evaluation

A possible way to prevent classification scores from looking unreasonably well is by applying cross-domain classification, i.e. testing a classifier on a dataset different from the one it was trained on. The specific biases we pointed out should be primarily restricted to individual datasets and not carry over to other ones. This is illustrated by Table 5. Compared to in-domain classification (Table 1), all classifiers perform worse. So all classifiers seem to be affected by data bias to some degree. Datasets with explicit abuse and less biased sampling (*Kaggle*, *Founta*, *Razavi*) still perform reasonably when trained among each other, i.e. they are not heavily affected, whereas datasets with implicit abuse and biased sampling (*Warner*, *Waseem*, *Kumar*) perform poorly. This time this also includes *Waseem* which implies that the good performance in in-domain classification (Table 1) was indeed caused by data bias.

Of course, cross-domain classification may not always be practical, particularly if a specific subtype of language abuse is studied for which there is only one dataset available. However, even then, simple methods such as computing the words that highly correlate with the different classes on that dataset, similar to what we did in Table 2, may already indicate that there are biases hidden in the dataset. If only a very small amount of biased words is identified, then usually it suffices to manually *debias* the dataset. By that, one understands sampling additional microposts containing the words manually detected to be biased (Dixon et al., 2017; Wiegand et al., 2018b). For example, in the case of the *Waseem*-dataset, randomly sampling additional tweets matching the words *announcer*, *commentator*, *football* or *sport*, would reduce the sexism bias we reported in this paper (simply because random tweets are unlikely to contain sexist remarks unlike the existing tweets from the *Waseem*-dataset).[5] In order to avoid author bias to interfere with classification, one could restrict the number of microposts per author. This would result in a more balanced distribution of microposts per author.

[5]Please note, however, that in the case of the *Waseem*-dataset, this form of debiasing would not completely solve the data bias since this dataset contains biased words beyond the four words mentioned above.

	test					
train.	**Kaggle**	**Founta**	**Razavi**	**Warner**	**Waseem**	**Kumar**
Kag.	N/A	**85.83**	**76.15**	63.91	60.32	62.48
Fou.	**84.70**	N/A	**70.11**	66.12	64.80	61.25
Raz.	**72.15**	**73.34**	N/A	60.22	61.76	60.61
War.	54.79	55.66	49.32	N/A	61.78	52.94
Was.	61.23	60.88	53.00	61.66	N/A	55.20
Kum.	69.31	65.93	62.98	55.74	59.20	N/A

Table 5: Cross-domain classification (*eval.: F1*).

7 Related Work

Previous work already established that identity terms (e.g. *gay*, *Jew* or *woman*) have a bias to co-occur with abusive language (Dixon et al., 2017; Park et al., 2018). In this work, we showed that this problem is not restricted to the small set of identity terms. Most biases are introduced by the sampling method used on a dataset and they have a huge impact on classification performance.

8 Conclusion

We examined the impact of data bias on abusive language detection and showed that this problem is closely related to how data have been sampled. On the popular *Waseem*-dataset, we illustrated that under more realistic settings, where such biases would be less prominent, classification performance is much lower than reported in research publications. Currently, datasets with a higher degree of implicit abuse are more affected by data bias. Such bias often goes unnoticed in in-domain classification which is why we recommend cross-domain classification. Our finding that under a realistic evaluation classification performance is actually quite poor particularly on implicit abuse, is also in line with assessments from industry on the quality of the state of the art[6] which suggests that there is still a long way to go.

Acknowledgements

The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References

Sweeta Agrawal and Amit Awekar. 2018. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In *Proceedings of the European Conference in Information Retrieval (ECIR)*. Grenoble, France, pages 141–153.

[6]https://www.businessinsider.de/facebook-ceo-zuckerberg-says-hate-speech-stumps-ai-2018-4

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the International Conference on World Wide Web (WWW)*. Perth, Australia, pages 759–760.

Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava, and Georg Rehm. 2017. Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication. In *Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL)*. Springer, Berlin, Germany, LNAI, pages 180–191.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2017. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. San Francisco, CA, USA.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behaviour. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. Stanford, CA, USA.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain, pages 427–431.

Mladen Karan and Jan Snajder. 2018. Cross-Domain Detection of Abusive Language Online. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 132–137.

Rohan Kshirsagar, Tyrus Cukuvac, Kathleeen McKeown, and Susan McGregor. 2018. Predictive Embeddings for Hate Speech Detection on Twitter. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 26–32.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC)*. Santa Fe, NM, USA, pages 1–11.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018a. Author Profiling for Abuse Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. Santa Fe, NM, USA, pages 1088–1098.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018b. Neural Character-based Composition Models for Abuse Detectiion. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 1–10.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pages 2799–2804.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. In *arXiv preprint arXiv: 1801.04433*.

Jing Qian, Mai ElSherief, Elizabeth M. Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. New Orleans, LA, USA, pages 118–123.

Amir Hossein Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*. Ottawa, Canada, pages 16–27.

Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. 2018. Learning Representations for Detecting Abusive Language. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 115–123.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 107–101.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proceedings of the Workshop on Abusive Language Online (ALW)*. Brussels, Belgium, pages 75–85.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Workshop on Language in Social Media (LSM)*. Montréal, Canada, pages 19–26.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*. Vancouver, BC, Canada, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop)*. San Diego, CA, USA, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a Lexicon of Abusive Words – A Feature-Based Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. New Orleans, LA, USA, pages 1046–1056.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval Workshop*. Vienna, Austria, pages 1–10.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International Conference on World Wide Web (WWW)*. Perth, Australia, pages 1391–1399.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. New York City, NY, USA, pages 3952–3958.