Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics

Hala Al Kuwatly*
TU Munich,
Department of Informatics,
Germany
hala.kuwatly@tum.de

Maximilian Wich*
TU Munich,
Department of Informatics,
Germany
maximilian.wich@tum.de

Georg Groh
TU Munich,
Department of Informatics,
Germany
grohg@in.tum.de

Abstract

Machine learning is recently used to detect hate speech and other forms of abusive language in online platforms. However, a notable weakness of machine learning models is their vulnerability to bias, which can impair their performance and fairness. One type is annotator bias caused by the subjective perception of the annotators. In this work, we investigate annotator bias using classification models trained on data from demographically distinct annotator groups. To do so, we sample balanced subsets of data that are labeled by demographically distinct annotators. We then train classifiers on these subsets, analyze their performances on similarly grouped test sets, and compare them statistically. Our findings show that the proposed approach successfully identifies bias and that demographic features, such as first language, age, and education, correlate with significant performance differences.

1 Introduction

According to the online harassment report published by Pew Research Center, "four-in-ten Americans have personally experienced online harassment, and 62% consider it a major issue." (Duggan, 2017, p.3). Online environments such as social media and discussion forums have created spaces for people to express their opinions and viewpoints, but this comes at the cost of hateful, offensive, and abusive content. Moderating this content manually requires a lot of staff and large amounts of hand-curated policies, which generated much interest in automatic content moderation systems that make use of recent advances in machine learning (Schmidt and Wiegand, 2017).

One challenge of training machine learning systems is the demand for large amounts of labeled data. Hence, many researchers use crowdsourcing platforms to annotate their data sets (Davidson et al., 2017; Founta et al., 2018; Vidgen and Derczynski, 2020), although having expert annotators has proven to improve the quality of annotations (Waseem, 2016). Such crowdsourcing approaches, however, exposes hate speech detection systems to annotator bias. Hateful behavior can take many forms (Waseem et al., 2017), making it harder to obtain a clean, common definition of hate speech, and resulting in subjective and biased annotations. Biases in the annotations are then absorbed and reinforced by the machine learning models, causing systematically unfair systems (Bender and Friedman, 2018). Therefore, it is not surprising that a large body of work has identified and mitigated this bias (Bender and Friedman, 2018; Bountouridis et al., 2019; Dixon et al., 2018).

We already know that people with particular demographic characteristics (e.g., black, disabled, or younger people) become more frequently targets of hate (Vidgen et al., 2019b). An aspect that is sparsely investigated in this context is the relation between annotators' demographic features and a potential bias in the data set. We want to fill this gap by addressing the following research question:

How do annotators' demographic features such as gender, age, education and first language impact their annotations of hateful content?

To answer this question, we conduct the following exploratory study: We sample balanced subsets of data that are labeled by demographically distinct annotators. We then train classifiers on these subsets, analyze their performances on similarly split test sets, and compare them statistically.

2 Related work

Since unintended bias in hate speech datasets can impair the model's performance (Waseem, 2016)

*These authors contributed equally to this work.

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 184–190

https://doi.org/10.18653/v1/P17

and fairness (Vidgen et al., 2019a; Dixon et al., 2018), a lot of recent work has been done to investigate this phenomenon (Wiegand et al., 2019; Kim et al., 2020).

Some work examined racial bias (Sap et al., 2019; Davidson et al., 2019; Xia et al., 2020), others explored gender bias (Gold and Zesch, 2018), aggregation bias (Balayn et al., 2018) and political bias (Wich et al., 2020b). The type of bias we are examining in this study is the annotator bias. Waseem (2016) studied the influence of annotator expertise on classification models and found that systems trained on expert annotations outperform those trained on amateur annotations, confirming and extending the results from Ross et al. (2017). Geva et al. (2019) showed that model performance improves when exposed to annotator identifiers, which suggests that annotator bias needs to be considered when creating hate speech models. Salminen et al. (2018) studied the difference between annotations of crowd workers from 50 countries and found those differences highly significant. Binns et al. (2017) examined the effect of the gender of the annotators on the performance of classifiers. Wich et al. (2020a) studied the similarities in the behaviour of the annotators to reveal biases that they bring into the data.

To the best of our knowledge, no one has developed a method to identify annotator bias based on multiple demographic characteristics of the annotators and measure its impact on the classification performance.

3 Data

We used the personal attack corpora from Wikipedia's Detox project (Wulczyn et al., 2017), which contains 115,864 labeled comments from Wikipedia on whether the comment contains a form of personal attack. The labels are the following (Wikimedia, n.d.):

- Quoting attack: Indicator for whether the annotator thought the comment is quoting or reporting a personal attack that originated in a different comment.
- Recipient attack: Indicator for whether the annotator thought the comment contains a personal attack directed at the recipient of the comment.
- Third party attack: Indicator for whether the annotator thought the comment contains a personal attack directed at a third party.
- Other attack: Indicator for whether the annotator thought the comment contains a personal attack but is not quoting attack, a recipient attack or third party attack.
- Attack: Indicator for whether the annotator thought the comment contains any form of personal attack. (Wikimedia, n.d.)

For our study, we used the attack label as the classification target label, not taking into consideration the other labels.

Feature	Trainset size	Testset size	Total size
Gender	4,401	1,100	5,501
First language	2,038	509	2,547
Age group	6,782	1,696	8,478
Education	3,174	794	3,968

Table 1: Number of comments in each demographic feature's datasets

The comments were labeled by 4,053 crowdworkers. For 2,190 of them, we have the demographic information. For each of these annotators we have the following demographic features:

- Gender: 'male' or 'female'
- English first language: '1' or '0'; '1' = annotator's first language is English
- Age group:'Under 18', '18-30', '30-45', '45-60', 'Over 60'. Since annotators are not equally distributed across age groups (see distribution plot in the appendix), we changed the grouping to 'Under 30' and 'Over 30'.
- Education (highest obtained education level): 'none', 'some', 'hs', 'bachelors', 'masters', 'doctorate', 'professional'. 'hs' is short for high school. Since annotators are not equally distributed across education levels (see distribution plot in the appendix), we changed the grouping to 'Below hs' (includes hs) and 'Above hs'.

4 Methodology

We address the research question by training classification models on data from demographically distinct groups and comparing their performances[1].

[1]Code available on GitHub: https://github.com/mawic/annotator-bias-demographic-characteristics

The hypothesis is that a statistically significant difference between the classifiers' performances indicates an annotator bias related to the studied demographic feature.

In the first step, we group the annotators by their demographic features, such as gender, age, education level, and native language. For each of those features, we create $m + 1$ datasets where m is the number of different values a demographic feature can take, e.g. for gender m could be equal to 2 if we only consider male and female annotators. All datasets have the same comments, but with different labels aggregated from annotators belonging to each different group. The additional dataset ($+1$) has labels aggregated from annotators belonging to all groups. It serves as a control group. We call this dataset the mixed dataset. We measured the inter-rater agreement within each group using Krippendorff's alpha (Hayes and Krippendorff, 2007).

In the second step, we split the datasets into train and test sets, and train 20 classifiers for each group on the group's training set and report F1 scores for all test sets. We train 20 classifiers to get multiple data points for each group's classifier and then apply the Kolmogorov-Smirnov test to examine whether they are significantly different [2]. The null hypothesis in this context is that the two samples are drawn from the same distribution. If we can reject the null hypothesis ($p < 0.05$) for a certain demographic feature, this will be evidence that annotators belonging to different groups of feature values hold different norms and are bringing in different biases into their annotations.

Concerning the classification model, we chose to make use of recent advancements in transfer learning and employ DistilBERT as a classifier due to the limited number of data points annotated by each group. DistilBERT (Sanh et al., 2019) is a smaller and faster distilled version of BERT (Devlin et al., 2018). In the context of abusive language detection, it provides a comparable performance (Vidgen et al., 2020). We used the base uncased version of DistilBERT (`distilbert-base-uncased`) with a maximum sequence length of 100, a learning rate of 5×10^{-6}, and 1cycle learning rate policy (Smith, 2018) and trained each classifier for 2 epochs.

[2]We trained 20 classifiers only for practical constraints.

4.1 Data split

To ensure the comparability of the classifiers, it is necessary to compile the training and test sets in the right way. Therefore, we define the following 2 conditions for selecting the comments: (1) All data sets of one feature contain the same comments. (2) At least 6 annotators from each demographic group annotated the comment. In the case of the gender group, that means a selected comment was annotated by at least 6 male and 6 female annotators.

For each demographic feature, we create 3 training and test set combinations. In the first one, the labels are taken from a random set of 6 annotators belonging to the first demographic group (e.g., males). In the second one, the labels of the comments are taken from a random set of 6 annotators belonging to the second demographic group (e.g., females). The third train and test sets are mixed: the labels of the comments are taken from a random set of 3 annotators belonging to the first demographic group and 3 annotators belonging to the second demographic group. While the subset of comments stays unchanged, for each of the 20 classifiers we sample the annotations of different random annotators. Data sets' sizes can be found in Table 1.

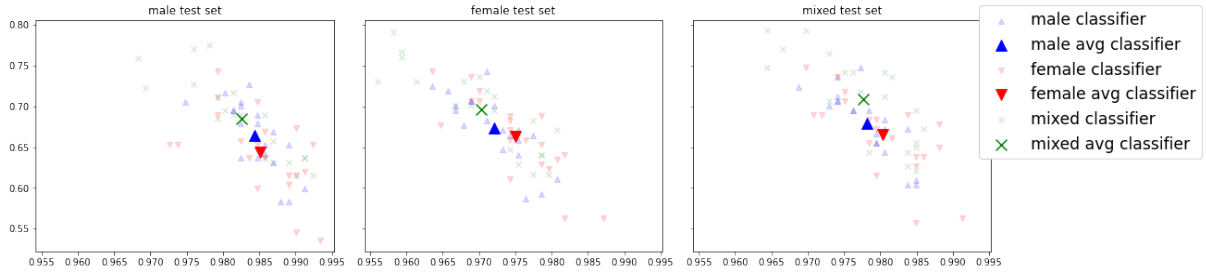
We also performed the same experiments without the limitation of sharing the same comments in the data sets of each feature, in order to increase the size of comments in the splits. Results were very similar to our shared comments experiments.

5 Results

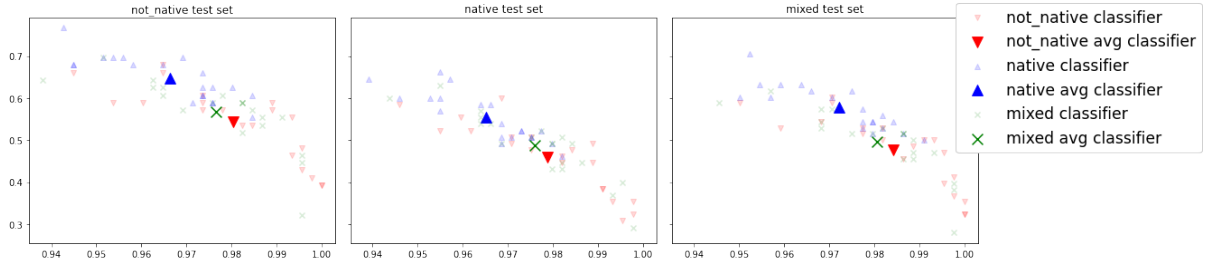
In this section, we report the results of our experiments for each demographic feature. The results comprise the inter-rater agreement of the annotators in the different groups, the averaged F1 scores of the trained classifiers, the sensitivity and specificity of the classifiers as charts, and the p-values generated by the Kolmogorov-Smirnov tests.

5.1 Gender

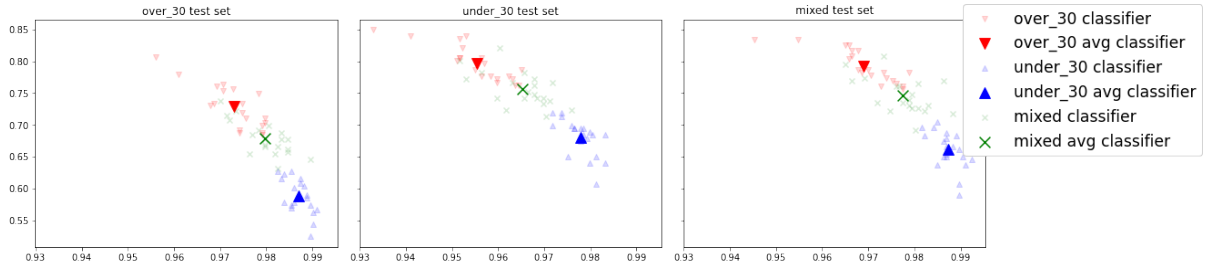
In regards to gender, we could not find evidence of any significant difference between male and female classifiers. Although the inter-rater agreement is significantly lower for females (0.45) than for males (0.51) (Table 4), the average F1 scores of the 20 classifiers trained for each group show no significant difference (Table 2). When analyzing the sensitivity and specificity graphs in Figure 1a,



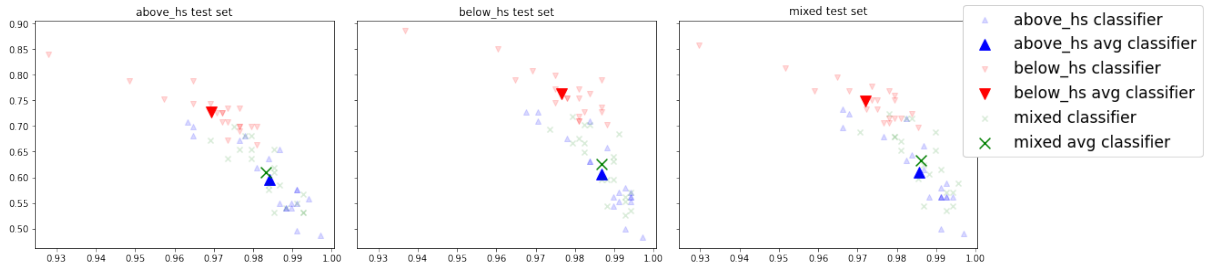
(a) Gender groups classifiers evaluated on gender groups test sets



(b) Language groups classifiers evaluated on language groups test sets



(c) Age groups classifiers evaluated on age groups test sets



(d) Education groups classifiers evaluated on education groups test sets

Figure 1: The x-axes are the specificity of the classifiers, and the y-axes are the sensitivity. Each transparent dot represents the specificity and sensitivity of each of the 20 classifiers trained for each group on the respective train set (dot marker) and evaluated on the respective test set (sub-figures). The opaque dots represent the average values.

one can also see no significant pattern or trend. The p-value resulting from the Kolmogorov-Smirnov test applied on the F1 scores of the 20 male classifiers and 20 female classifiers evaluated on the mixed test set is 0.83 (Table 3). Since it is larger than 0.05, we cannot conclude that a significant difference between the male and female classifier exists.

5.2 First Language

Our experiments on first language classifiers resulted in the following observations:

1. Classifiers trained on native-labeled data have a notably higher F1 score (Table 2) and are also more sensitive to all test sets (the blue triangles in Figure 1b), which suggests that they are particularly better at classifying comments

trainset \ testset	male	female	mixed
male	0.850	0.855	0.829
female	0.846	0.859	0.838
mixed	**0.856**	**0.862**	**0.848**
	native	not native	mixed
native	**0.814**	**0.818**	**0.816**
not native	0.768	0.786	0.764
mixed	0.783	0.778	0.772
	under 30	over 30	mixed
under 30	0.853	0.833	0.863
over 30	0.858	**0.870**	**0.883**
mixed	**0.860**	0.860	0.879
	below hs	above hs	mixed
below hs	**0.885**	**0.861**	**0.873**
above hs	0.839	0.830	0.839
mixed	0.847	0.836	0.850

Table 2: Average F1 scores of the classifiers.

Feature	p-value
Gender	8.3×10^{-1}
First Language	1.0×10^{-3}
Age group	1.1×10^{-8}
Education	1.4×10^{-7}

Table 3: Results of the Kolmogorov-Smirnov test, inputs to the tests are the F1 scores of the 20 classifiers evaluated on the mixed test set of each feature.

that contain personal attack.

2. Classifiers trained on only non-native-labeled data perform almost as good as the baseline (classifier trained on mix-labeled data) (Table 2).

3. We found very minor disparities in the specificity of both classifiers (Figure 1b).

The result of the Kolmogorov-Smirnov test on native and non-native classifiers is a p-value of 1.0×10^{-3} (Table 3), thus we can reject the null hypothesis and conclude that a significant difference does exist between them.

5.3 Age group

Our experiments resulted in the following observations:

1. Classifiers trained on over-30-labeled data have higher F1 scores than classifiers trained on under-30 labeled data on all test sets. They are however comparable to the baseline (classifier trained on mix-labeled data) (Table 2).

2. All classifiers are less sensitive to over-30-labeled test set (Figure 1c), which might suggest that it contains harder examples that all classifiers failed to correctly classify.

Feature	Group	Inter-rater Agreement
Gender	Male	0.51
	Female	0.45
	Mixed	0.48
English	Native	0.46
	Not native	0.50
	Mixed	0.48
Age group	Under 30	0.47
	Over 30	0.50
	Mixed	0.48
Education	Below hs	0.49
	Above hs	0.48
	Mixed	0.48

Table 4: Inter-rater agreement for all groups

The Kolmogorov-Smirnov test on the results of the two classifiers produces a p-value of 1.1×10^{-8} (Table 3), thus we can reject that they come from the same distribution and conclude that a significant difference does exist between them.

5.4 Education

Our experiments resulted in the following observations:

1. The F1 scores of the classifiers trained on below-hs-labeled data are higher than scores of classifiers trained on above-hs-labeled data on all test sets (Table 2).

2. Classifiers trained on below-hs-labeled data have a comparable specificity to the other classifiers but with a notably higher sensitivity on all test sets. (Figure 1d).

The Kolmogorov-Smirnov test with a p-value of 1.4×10^{-7} (Table 3) also shows that there exists a significant difference between the two groups.

6 Discussion

In light of our results, we can conclude that the gender of the annotator does not bring a significant bias in annotating personal attacks in the studied dataset. However, when Binns et al. (2017) explored the role of gender in *offensive* content annotations, they established a distinguishable difference between males and females. We think this is related to the nature of the annotation task itself. To investigate other tasks, our approach can further be applied in future work on the other data sets provided by Wikipedia's Detox project (Wulczyn et al., 2017) such as aggressiveness and toxicity to investigate the effects of gender for those tasks.

When it comes to the first language of the annotators, it seems that native English speakers are gen-

erally better at identifying personal attacks in comments. The results also suggest that non-natives could not capture attack in comments that natives found to contain attack.

In addition, age groups and education levels of the annotators also seem to play a notable role in how attacks are perceived. Training a classifier on aggregated labels from all groups, even if the data is balanced between groups, does not seem to be fair to all groups involved.

Although we have only explored the demographic features provided by the data set and grouped some of them for reasons dictated by the data size, we think other features (e.g., race, ethnicity, and political orientation), different within feature groupings and feature intersections might produce new biases. While exploring all possible demographic features prior to building models is simply infeasible, the set of studied features can be determined per task.

Our approach demonstrated how particular training sets labeled by different groups of people can be used to identify and measure bias in data sets. These biases are never constant or static even within one group, for what counts as hateful is always subjective. In consequence, having only one version of ground truth is bound to produce biased systems. It is inevitable that training models on biased datasets produces systems that amplify those biases, whether these biases are exclusionary, prejudicial, or historical. Therefore and due to the conflicting and ever-changing definitions of hate speech among communities, we urge researchers in the hate speech domain to examine their data sets closely and thoroughly in order to understand their limitations and consequences.

7 Conclusion

This work explored bias in hate speech classification models where the task is inherently controversial and annotators' demographic data might influence the labels. We demonstrate how particular demographic features might bias the models in ways that are important to look into prior to using such models in production. We explored the performance of classification models trained and tested on different training and test data splits, in order to identify the fairness of these classifiers and the biases they absorb. We hope that our proposed method for identifying and measuring annotator bias based on annotators' demographic characteristics will help to build fairer hate speech classifiers.

Acknowledgments

This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

References

Agathe Balayn, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, and Zoltán Szlávik. 2018. Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*, volume 2276. CEUR.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.

Dimitrios Bountouridis, Mykola Makhortykh, Emily Sullivan, Jaron Harambam, Nava Tintarev, and Claudia Hauff. 2019. Annotating credibility: Identifying and mitigating bias in credibility datasets.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Maeve Duggan. 2017. *Online Harassment 2017*. Pew Research Center.

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 behavior. *arXiv preprint arXiv:1802.00393*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Michael Wojatzki Tobias Horsmann Darina Gold and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10.

Leslie N Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting east asian prejudice on social media.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. How much online abuse is there? a systematic review of evidence for the uk. *The Alan Turing Institute*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proc. 1st Workshop on NLP and Computational Social Science*, pages 138–142.

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

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020a. Investigating annotator bias with a graph-based approach. In *Proc. 4th Workshop on Online Abuse and Harms*.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020b. Impact of politically biased data on hate speech classification. In *Proc. 4th Workshop on Online Abuse and Harms*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

Wikimedia. n.d. Research:detox/data release. https://meta.wikimedia.org/wiki/Research:Detox/Data_Release.

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

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*.