Investigating Annotator Bias in Abusive Language Datasets

Maximilian Wich, Christian Widmer, Gerhard Hagerer, Georg Groh
Technical University of Munich, Germany
{maximilian.wich,christian.widmer,gerhard.hagerer}@tum.de,
grohg@in.tum.de

Abstract

Nowadays, social media platforms use classification models to cope with hate speech and abusive language. The problem of these models is their vulnerability to bias. A prevalent form of bias in hate speech and abusive language datasets is annotator bias caused by the annotators subjective perception and the complexity of the annotation task. In our paper, we develop a set of methods to measure annotator bias in abusive language datasets and to identify different perspectives on abusive language. We apply these methods to four different abusive language datasets. Our proposed approach supports annotation processes of such datasets and future research addressing different perspectives on the perception of abusive language.

1 Introduction

A challenge that social media platforms are facing in recent years is the large amount of hate speech and other forms of abusive language (Duggan, 2017). Manual monitoring, however, is no longer possible due to the vast volume of user-generated content. Therefore, machine learning models are trained and used by social media platforms, such as Facebook, to automatically detect such content (Kantor, 2020). According to Rose (2021), these models are a key component of Facebook's fight against hate speech.

A problem with such machine learning models is that they are vulnerable to bias (Vidgen and Derczynski, 2021; Dixon et al., 2018). Biased models can strongly impair the fairness of a system, which can lead to discrimination (Dixon et al., 2018).

Bias in abusive language detection is already a topic that researchers have started to investigate (Vidgen and Derczynski, 2021; Wich et al., 2021). The type of bias we will focus on in this study is annotator bias. This form of bias is a result of annotators who perceive abusive language differently from each other and have different levels of experience and knowledge (Ross et al., 2016; Waseem, 2016; Geva et al., 2019; Wich et al., 2020).

We aim to investigate two aspects of annotator bias. (1) Assuming that there is only one perspective (one truth) on whether a text is abusive or not, we develop an approach to measure and visualize annotator bias. This approach optimizes the annotation process (e.g., outlier detection, adapting appropriate annotation guidelines). (2) Acknowledging multiple valid views on a text (e.g., a group has a more liberal attitude towards abusive texts, while another is stricter), we aim to identify annotator groups to model different, yet valid perspectives. The questions resulting from these research objectives are the following:

- RQ1: How can we measure and visualize annotator bias in abusive language datasets?
- RQ2: How can we identify and visualize different annotator perspectives on abusive language?

Our contributions are the following:

1. To characterize annotators, we gauge how liberal or strict they annotate in comparison to other annotators. To model annotator bias, we calculate a pessimistic and optimistic score for each annotator that can be visualized in different ways (e.g., scatter plot, cluster map). We apply it to four abusive English language datasets with different groups of annotators.
2. To identify annotator groups with different annotator groups with different perspectives on abusive language, we utilize a classifier-based method with the proposed approach, which is applied to one dataset.

Proceedings of Recent Advances in Natural Language Processing, pages 1515–1525
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_170

Name	Documents	Source	Labels	Annotators	Expert check	Reference
Vidgen	20,000	Twitter	**hostility**, criticism, counter speech, discussion of east Asian prejudice, non-related	26	yes	Vidgen et al. (2020)
Guest	6,567	Reddit	**misogynistic**, non-misogynistic	6	yes	Guest et al. (2021)
Kurrek	40,000	Reddit	**derogatory usage**, appropriative usage, non-derogatory usage, homonyms	20	yes	Kurrek et al. (2020)
Wulczyn	115,864	Wikipedia (discussion)	**attack**, non-attack	4,053	no	Wulczyn et al. (2017)

Table 1: Overview of selected abusive datasets (class names in bold are considered as abusive, the others as neutral).

2 Related Work

Hate speech and abusive language detection have gained a lot of attention in recent years. A range of different datasets (Vidgen and Derczynski, 2021) and shared tasks (Basile et al., 2019; Zampieri et al., 2019, 2020) were published to foster research in this area. Most of the datasets are commonly labeled by crowdworkers or those in academia with varying expertise (Vidgen and Derczynski, 2021). However, human annotations tend to be subjective and thus inconsistent (Aroyo and Welty, 2015), at least if not moderated very strictly. Especially for abusive language, Salminen et al. (2018) show that individuals interpret hate speech differently. One common method to improve the label quality is presenting each sample to multiple annotators and aggregate their results (Sheng et al., 2008). Dawid and Skene (1979) were the first to propose an approach that incorporates annotator quality into label aggregation. Their expectation-maximization (EM) algorithm uses the bias matrices to estimate the latent truth. In the matrices the annotator quality is encoded. Their seminal work led to further improvements and methods (Whitehill et al., 2009; Raykar and Yu, 2012; Hovy et al., 2013). For NLP tasks, Snow et al. (2008) used a customized Dawid-Skene algorithm to correct for individual biases of crowdworkers and improve model accuracy. However, they did not quantify and inspect the bias of the annotators.

In abusive language detection, annotator bias research has focused on how the annotators background influences their annotations. Waseem (2016) found models trained on crowd annotations are outperformed by models trained on expert annotations. Ross et al. (2016) emphasized the importance of detailed guidelines to achieve reliable annotations. Binns et al. (2017) showed that classifiers trained on annotations differ in their performance on test data annotated by men or women. Al Kuwatly et al. (2020) picked up this approach, enhanced it, included other demographics, and found significant differences for annotator's age group and educational background. Sap et al. (2019) observed that posts in African American dialect are more likely to be labeled offensive. Similarly, Larimore et al. (2021) found that white and non-white workers annotate racially sensitive topics differently. Apart from studying the demographic background, researchers also attempt to find groups of annotators with common annotation behavior. Wich et al. (2020) use graph methods to cluster annotators in groups with higher inter-annotator agreement within groups than across groups. Akhtar et al. (2020) defined a polarization measure to split annotators in two groups that maximize opposing annotations. To the best of our knowledge, no one has quantified annotator bias at the annotator level. Furthermore, the hypothesis of multiple perspectives on abusive language is rarely investigated.

3 Datasets

We use four different abusive English language datasets to demonstrate our proposed approach. It was challenging to find appropriate datasets because our experiment requires unaggregated annotation data. Most of the abusive language datasets contain only the agreed upon labels in the documentation and not the individual votes of the annotators.

Table 1 lists the four datasets with additional in-

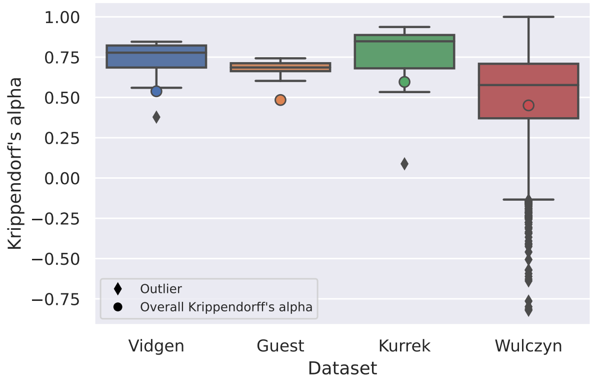
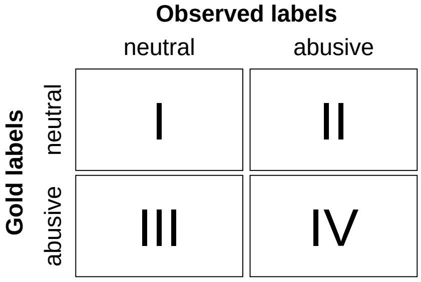


Figure 1: Box plots of the annotators' inter-rater reliability scores.



		Observed labels	
		neutral	abusive
Gold labels	neutral	I	II
	abusive	III	IV

Figure 2: Bias matrix of an annotator.

formation. The first three datasets (*Vidgen*, *Guest*, and *Kurrek*) are similar because they are annotated by small groups of annotators (between 6 and 26). Furthermore, each document of the three datasets was annotated by two annotators. In case of disagreement, an expert reviewed the votes and decided on the gold label. In contrast, the *Wulczyn* dataset was annotated by many crowdworkers—a typical crowdsourcing setup: a group of workers who annotated a small number of documents. Each document was annotated by up to 10 annotators. In case of ambiguous annotations, an expert review did not take place. The gold label was determined based on majority vote. For our experiment, we convert all datasets to a binary task (abusive/neutral) to compare the results.

Figure 1 shows the distributions of the annotators' inter-rater reliability scores in form of Krippendorff's alpha. The colored dots represent the overall inter-rater reliability score of each dataset. We see that the overall Krippendorff's alphas are all in the same range. The *Wulczyn* dataset, however, exhibits a considerable variance in contrast to the other three datasets. The reason is that 4,053 crowdworkers annotated this dataset, while an instructed small group of workers annotated the other three datasets. Therefore, we see many outliers. In the case of the *Vidgen* and *Kurrek* dataset, only one annotator strongly differs from the others.

4 Methodology

Our analysis of the annotator bias in the selected abusive language datasets consists of two parts. In the first part, we characterize the annotation behavior based on the deviations of the annotator votes compared with the gold standard of the dataset. In the second part, we visualize the perspectives of different annotator groups on abusive language with the aid of classification models.

4.1 Characterizing Annotator Bias

We define annotator bias as the deviations between the annotator votes and the gold labels of the dataset. The gold labels are either the final labels of the dataset or majority of the single votes. To measure the annotator bias, we use the concept of the confusion matrix. Figure 2 shows a matrix for a binary classification task of abusive documents (neutral/abusive). The rows represent the classes of the gold labels; the columns represent the classes observed by the annotator. The bias matrix quantifies the deviations between the labels observed by the annotator and the gold labels. Each annotator has one bias matrix.

We use cells II and III, which represent false negatives (type II error) and false positives (type I error) in the classical confusion matrix, to characterize the annotators' behavior, and we assign each annotator a pessimistic and optimistic score. Cell II reflects the number of documents that are neutral according to the gold standard but that are annotated as abusive by the annotator, signaling that the annotator is pessimistic in these cases. Cell III is the opposite, and shows the number of documents that are labeled as abusive according to the gold standard but perceived as neutral by the annotator, signaling that the annotator is optimistic in these cases. The pessimistic (p_i) and optimistic (o_i) scores of an annotator (i) are entries II and III of row-normalized bias matrix. The concept of annotator's optimism and pessimism was proposed by Dawid and Skene (1979). This method also works if we have more than two classes as long as they are ordinal. In this case, the cells above or below the diagonal are summed up. In our paper, however, we consider only binary classification of tasks.

To analyze bias matrices, we utilize these options:

1. We calculate the bias matrix for a group of annotators or all of them by averaging the bias matrices. The resulting bias matrix delineates whether the selected annotators tend to be optimistic or pessimistic. These findings assist with adjustment of annotation guidelines or the training of the annotators.

2. We utilize a 2-dimensional scatter plot of the pessimistic and optimistic scores to visualize the annotators and their biases. In contrast to comparison of inter-rater reliability scores, this visualization reveals whether annotators are more optimistic or pessimistic than the gold standard. Such information can help to detect outliers in the respective direction and to instruct the identified annotators as appropriate.

3. We can calculate a distance between two annotators based on their bias matrices (A and B) with the Frobenius norm (Golub and Van Loan, 2013, p.71):

   $$distance(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}-b_{ij})^2}$$

   Visualizing these distances with a hierarchically clustered heatmap helps identify annotator groups with similar annotation behavior and outliers.

4. If the number of annotators is so large that the results of the previously proposed methods is no longer manageable, we can apply a hierarchical clustering on the bias matrices based on our distance metric. By doing so, annotators with a similar annotation behavior are clustered. If we aggregate the bias matrices according to (1), we observe how the cluster annotated the data in context of the gold standard.

5. If we have additional information about the annotators to characterize them (e.g., demographics, such as age or education), we can use the pessimistic and optimistic scores to test whether there is a significant difference between the annotation behavior of annotators with different characteristics. For this purpose, we apply the two-dimensional Kolmogorov-Smirnov (KS2D) test (Fasano and Franceschini, 1987; Peacock, 1983) to compare the distributions of the groups' pessimistic and optimistic scores. The output of the test is the Kolmogorov-Smirnov statistic D and the corresponding significance level s. If D is larger than the predefined significance level p and p is larger than s, we can reject the null hypothesis that both samples have the same distribution. We use the Python implementation provided by Gabriel Taillon[1]. In the case of the *Wulczyn* dataset, we have such information (Wulczyn et al., 2017). Our predefined significance level p is 0.05.

4.2 Identifying Different Perspectives on Abusive Language

The previous subsection focuses on methods to measure and visualize annotator bias, answering RQ1. The underlying assumption is that there is one truth, meaning one valid perspective on abusive language, and we want to identify outliers deviating from the one truth.

Now we assume that there are more perspectives on abusive language—e.g., a group has a more liberal attitude toward abusive texts, while another group is less liberal. To examine this hypothesis, we run the following experiment. First, we split the annotators into different groups based on the pessimistic and optimistic scores. Second, for each group we create a dataset, containing the documents that all groups annotated. The labels of the documents result from the majority vote of the groups' annotators. Third, for each group we train a classification model on its training set and evaluate it on the test sets of all groups. Suppose a classifier performs well on its test set and worse on the other test sets. Thus, the performance is comparable to a baseline classifier trained on the same data with gold labels. In that case, it indicates that this group has a coherent perspective on abusive language that differs from the other groups. This approach is based on the method proposed by Wich et al. (2020).

To split the annotators according to their pessimistic (p_i) and optimistic o_i scores, we apply the following function:

$$group_a(p_i, o_i) = \begin{cases} 0 & \text{if } p_i \geq 3 \cdot o_i \\ 1 & \text{if } o_i \geq 3 \cdot p_i \\ 2 & \text{otherwise} \end{cases}$$

The factor 3 in the function is the result of a trade-off between having a dominating dimension

[1] https://github.com/Gabinou/2DKS

in the optimistic and pessimistic group and having enough annotators in all three groups. Increasing the factor would strengthen the dominating dimensions but reduce the number of annotators in the optimistic and pessimistic groups. Decreasing the factor would weaken the dominating dimension but increase the number of annotators in the groups.

For the classification model, we use the pre-trained English DistilBERT model `distilbert-base-uncased` provided by the Transformer Library from Hugging Face (Wolf et al., 2020); it is a more concise version of BERT (Sanh et al., 2019; Devlin et al., 2019) and provides a performance comparable to BERT for abusive language detection (Devlin et al., 2019). We train each model for three epochs with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 32. After the three epochs, we select the model with the lowest validation loss. 60% of the documents annotated by all groups are used as a training set, 20% as a validation set, and 20% as a test set. To compare the different classifiers, we use the macro F1 score.

5 Results

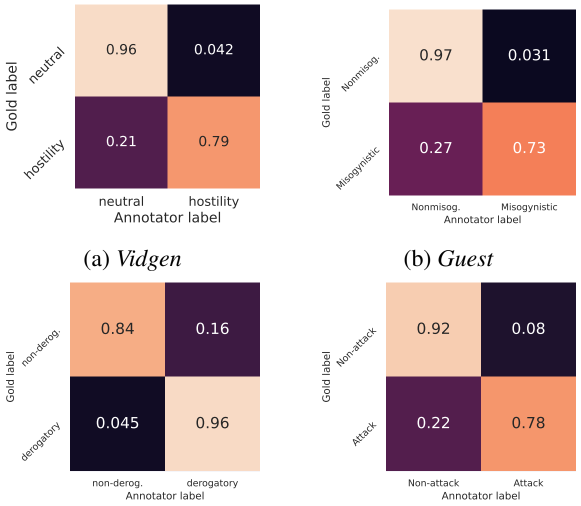
5.1 Characterizing Annotator Bias

Aggregated Bias Matrix

The problem of the inter-rater reliability analysis is that it does not reveal whether the annotators annotated more pessimistically or optimistically. This gap is addressed by the aggregated bias matrices, shown in Figure 3. The annotators of the datasets *Vidgen*, *Guest*, and *Wulczyn* tended to annotate more liberally because the optimistic scores (bottom-left cell) outweigh the pessimistic scores (upper-right cell). On the contrary, the annotators of the *Kurrek* dataset were stricter because 16% of non-derogatory documents were labeled as derogatory (pessimistic score), while only 4.5% (optimistic score) of the derogatory documents were labeled as non-derogatory.

Scatter Plot of Annotators

To gain a better understanding of the individual annotation behavior, we analyze the annotators based on their pessimistic and optimistic scores, shown in Figure 4. Considering the plots of *Vidgen*, *Guest*, and *Kurrek*, we observe that the annotators of *Vidgen* and *Guest* annotated more liberally due to the higher optimistic scores, while it is the opposite for the Kurrek dataset. Comparing the *Guest* dataset with the other two, we see that the annotators are less widely spread, meaning the annotation behavior is more coherent. Concerning the previously mentioned outliers of *Vidgen* and *Kurrek*, we can use the plots to better understand how they deviate. The outlier of *Vidgen* is the most right data point, the outlier of *Kurrek* is the uppermost data point. Their positions reveal that the outlier of *Vidgen* annotated more liberally, while the outlier of *Kurrek* was stricter. These findings can help instructors to guide the annotators if the method is used during the annotation process. A further observation concerning both datasets is that the density of annotators increases toward the origin of both dimensions. This indicates that most of the annotators have a similar annotation behavior.



(a) *Vidgen*

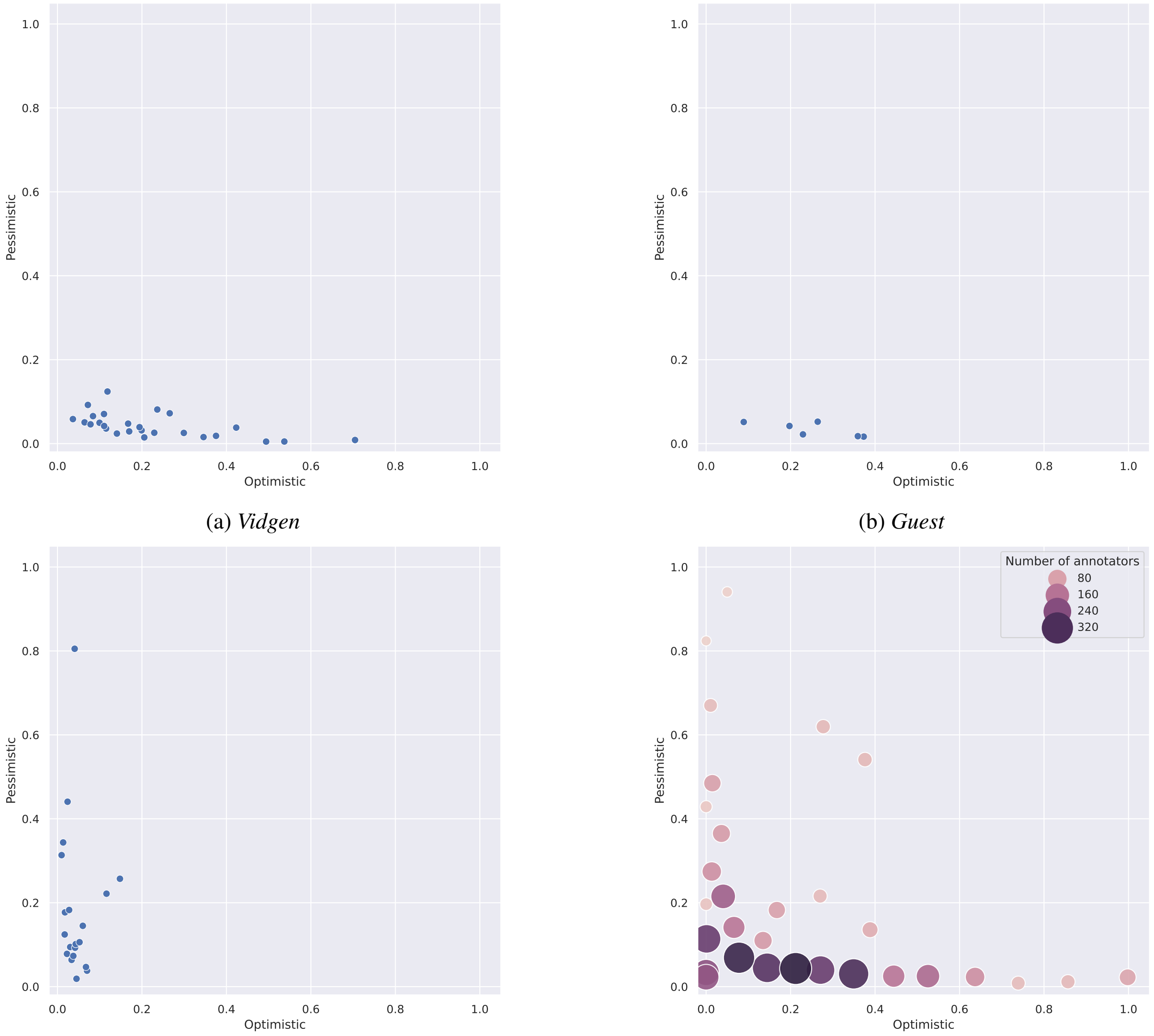
(b) *Guest*

(c) *Kurrek*

(d) *Wulczyn*

Figure 3: Aggregated bias matrices for the selected datasets.

In the case of the *Wulczyn* dataset, plotting each annotator as a data point would be confusing because the dataset contains 4,053 annotators. Therefore, we decided to cluster the annotators with a hierarchical clustering approach, facilitating data interpretation. We chose the agglomerative clustering approach with $k = 30$ and Euclidean distance function. The reason for $k = 30$ is that it is a manageable amount of data points on the scatter plot and it has the same order of magnitude as *Viden* and *Kurek*. Figure 4d shows the annotators' clusters. We observe the tendency of the annotators to annotate more liberally, as shown by the aggregated bias matrix in Figure 3d.



(a) *Vidgen*

(b) *Guest*

(c) *Kurrek*

(d) *Wulczyn* (clustered annotators)

Figure 4: Annotators visualized based on their pessimistic and optimistic scores; in case of *Wulczyn*, annotators are hierarchically clustered.

Cluster Map of Distances between Annotators

A method to identify groups of annotators with similar annotation behavior is the hierarchically clustered heatmap based on the distances between the bias matrices of the annotators. Figure 5 shows the cluster map of the *Kurrek* dataset. The first thing that catches the reader's eye is the first column and row. It shows the outlier of the dataset. Furthermore, we observe that the annotators Ann7, Ann13, Ann15, Ann3, and Ann5 (last five columns and rows) form a group. In Figure 4c, these annotators are the points above a pessimistic score of 0.2 and below 0.6. The other 15 annotators exhibit a more coherent annotation behavior. Due to the page restriction, we do not include the analysis for the other three datasets.

Different Annotation Behavior of Demographic Groups

The *Wulczyn* dataset contains demographic information for 2,190 of the 4,053 annotators (i.e., gender, age group, education, and English as the first language). We tested for each demographic feature whether there is a difference between the groups regarding the annotators' pessimistic and optimistic scores. The result of the two-dimensional Kolmogorov-Smirnov test for the demographic feature of gender is the following:

$$D_{\text{gender}} = 0.092 \qquad s_{\text{gender}} = 0.005$$

Based on this result, we can reject the null hypothesis ($p = 0.05$). Consequently, there is a significant difference between the pessimistic and optimistic

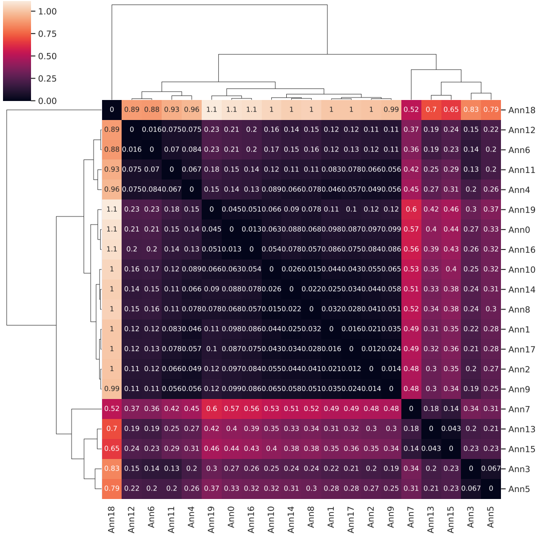


Figure 5: Cluster map of annotators' distances from *Kurrek* dataset.

scores of male and female annotators. Females are more pessimistic than males ($p_{\text{female}} = 0.107$ and $p_{\text{male}} = 0.090$), while the optimistic scores are comparable ($o_{\text{female}} = 0.192$ and $o_{\text{male}} = 0.199$). For the feature describing whether English is the first language of the annotator or not, we can also reject the null hypothesis:

$$D_{\text{1st language}} = 0.192 \qquad s_{\text{1st language}} = 8.9{\times}10^{-8}$$

Native English speakers have a larger pessimistic score ($p_{\text{native}} = 0.093$ and $p_{\text{non-native}} = 0.117$) and a lower optimistic score than non-native speakers ($o_{\text{native}} = 0.160$ and $o_{\text{non-native}} = 0.204$).

Table 3a shows the output of the two-dimensional Kolmogorov-Smirnov test for the different age groups. We observe that there are significant differences in the distributions of the annotators' pessimistic and optimistic scores between the age groups—except between the ages 30-45 and over 60 and 45-60 and over 60. Interestingly, the largest difference is between the age groups 18-30 and 45-60. While annotators between 45 and 60 are more pessimistic ($p_{45-60} = 0.146$ and $o_{45-60} = 0.128$), it is the opposite for annotators between 18 and 30 ($p_{18-30} = 0.08518$ and $o_{18-30} = 0.234$).

Table 3b shows the output for the different educational backgrounds. In contrast to the age groups, the scores of the annotators do not differ greatly between the groups; however, the difference between annotators who have a Bachelors and Masters degree is significant.

5.2 Identifying Different Perspectives on Abusive Language

Since this experiment requires a dataset with a large number of documents and annotators, we could conduct it only with the *Wulczyn* dataset. In the case of the other three datasets, the number of annotators is too small to meaningfully split the annotators into subsets and to have enough documents that were annotated by all subsets.

The results of the experiment to identify different perspectives and to answer RQ2 can be found in Table 2. It shows the different F1 scores for the abusive class of the classifiers that were trained on subsets of annotators (rows) and were evaluated on the test sets of these subgroups (columns).

	Pessimistic	Medium	Optimistic
Pessimistic	80.2	80.6	71.0
Medium	73.5	81.9	83.1
Optimistic	64.3	74.4	87.5

Table 2: F1 scores from classifiers of the different annotator subsets.

To answer our RQ2 on how to identify and visualize different perspectives on abusive language of the annotators, we need to focus on the pessimistic and optimistic data. We observe that the classifier trained on the annotations of the optimistic annotators performs best on its own test set (87.5%) and worst on the pessimistic test set (64.5%). When the classifier trained on the more pessimistic annotations, the result is the opposite. It performs most poorly on the optimistic test set (71.0%) and comparable well on its own test set (80.2%). Only on the test set of the medium group, the pessimistic classifier performs slightly better.

It is more relevant to our research question that the pessimistic and optimistic classifiers work well on their own test set but worse on the test set of the other extreme. The first fact indicates that the annotations are coherent, so that the classifier can learn patterns to identify abusive language. The second aspect shows that the labels of the pessimistic and optimistic subgroups' dataset are so different that it can cause a difference of 9.2 or 23.2pp in the F1 score. Consequently, we conclude that the annotators of the pessimistic and optimistic subgroup have two different perspectives on abusive language.

An explanation for the more coherent results of the optimistic classifier can be the larger number

	Under 18	18-30	30-45	45-60	Pessimistic	Optimistic
Under 18	-	-	-	-	0.080	0.172
18-30	**0.261 / 0.040**	-	-	-	0.085	0.234
30-45	**0.303 / 0.011**	**0.177 / 0.000**	-	-	0.100	0.165
45-60	**0.435 / 0.000**	**0.322 / 0.000**	**0.216 / 0.000**	-	0.146	0.126
Over 60	**0.416 / 0.031**	**0.377 / 0.016**	0.248 / 0.249	0.165 / 0.775	0.125	0.140

(a) Age group of annotators

	some	hs	bachelors	masters	doctorate	Pessimistic	Optimistic
some	-	-	-	-	-	0.085	0.210
hs	0.116 / 0.738	-	-	-	-	0.096	0.195
bachelors	0.109 / 0.790	0.059 / 0.341	-	-	-	0.096	0.193
masters	0.141 / 0.520	0.070 / 0.378	**0.102 / 0.040**	-	-	0.098	0.206
doctorate	0.175 / 0.827	0.217 / 0.407	0.199 / 0.516	0.231 / 0.346	-	0.075	0.216
professional	0.136 / 0.597	0.104 / 0.134	0.070 / 0.530	0.124 / 0.074	0.184 / 0.647	0.109	0.190

(b) Educational background of annotators

Table 3: Results of 2-dimensional Kolmogorov-Smirnov test for split according to demographic features and corresponding pessimistic and optimistic scores (*Wulczyn* dataset); first number in cells is D, second s; bold means rejected.

of annotators. While it comprises annotations from 1,708 annotators, the pessimistic subset contains only 1,312. As we can see, this difference is in line with the finding that the annotators of the *Wulczyn* dataset tended to annotate more liberally.

6 Discussion

The first part of our study addressing RQ1 shows that the proposed approach based on the pessimistic and optimistic scores helps to measure and visualize the difference in the annotation behavior of annotators. In contrast to the inter-rater reliability, our method reveals information about the tendency of the annotators: Did they annotate more liberally or stricter than the group average? These findings can be used to understand outliers better, instruct single annotators in the right direction to align them with the rest of the group and/or adapt the annotation guidelines. Our approach comprises a range of methods, from an aggregated perspective on all annotations to cluster analyses to evaluations of individual annotators. This variety allows handling of datasets with different numbers of annotators.

The proposed approach is unsupervised by itself because it does not require any labeled data. But it can be combined with additional data, as shown by the experiment with the demographic features. We showed that it can help to detect annotator bias caused by different demographic backgrounds. Our results are partially in line with the findings from Al Kuwatly et al. (2020), who examined the same dataset but with a different approach. We confirmed the differences between native and non-native speakers and between the age groups. In our case, we identified a significant difference between male and female annotators, which Al Kuwatly et al. (2020) did not find. In contrast to our experiment, they observed a greater difference between educational backgrounds. The reason for the discrepancy can be the different methods. They trained classifiers on different subsets and compared their performances, as we did for the second part of our study. Furthermore, they had to group the educational backgrounds to have enough data. Consequently, the results can differ. The advantage of our approach over the classifier-based method used by Al Kuwatly et al. (2020) and by Binns et al. (2017) on another dataset is that we do not rely on a classifier as we can use the full dataset.

The underlying assumption for the first part of the study is that there is only one foundational truth whether a text is abusive or not to demonstrate that we all share the same understanding. In the second part of the study, we had the controversial assumption that there are different perspectives on the perception of abusive language. Our goal was to use our proposed method to identify different perspectives and to visualize the differences. By splitting the annotators according to the ratio between the pessimistic and optimistic scores and training different classifiers for these annotators subsets, we showed that there are different perspectives on abusive language. The classifiers of the pessimistic and optimistic annotator subsets perform well on their own test set and poorly on the test set of the other subset. That means that the perception of abusive language within each group is coherent, but it differs from the perception of the other subset.

Multiple perspectives on abusive language should be further investigated. Akhtar et al. (2020), for example, showed that balancing different perspectives in the training set can improve the classification performance. We can also imagine building classification models that demonstrate different perspectives; each group would have a customized model based on the groups' individual values and perceptions.

7 Conclusion

In this paper, we presented a novel approach to measure and visualize annotator bias purely on their annotation behavior. This approach fosters a better understanding of annotation behavior, detecting outliers, and gaining insights on how to adapt annotation guidelines. Furthermore, we showed that there can be different perspectives on abusive language. Using our proposed approach, we can identify these perspectives and examine the differences.

Resources

The code is available under `https://github.com/mawic/annotator-bias-abusive-language`.

Acknowledgments

This paper is partially based on a joined work in the context of Christian Widmer's master's thesis (Widmer, 2021). This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research. The paper is based upon work partially supported by Google Cloud Platform research credits.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

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.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer errorrates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Giovanni Fasano and Alberto Franceschini. 1987. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170.

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. In *2019 Conference on Empirical Methods in Natural Language Processing*, pages 1161–1166.

Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*, volume 4. The Johns Hopkins University Press.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Arcadiy Kantor. 2020. Measuring Our Progress Combating Hate Speech .

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

John A Peacock. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627.

Vikas C Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 13(1):491–518.

Guy Rose. 2021. Community Standards Enforcement Report, First Quarter 2021.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.

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, Florence, Italy. Association for Computational Linguistics.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Rion Snow, Brendan Oconnor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Zeerak Waseem. 2016. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22:2035–2043.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Maximilian Wich, Tobias Eder, Hala Al Kuwatly, and Georg Groh. 2021. Bias and comparison framework for abusive language datasets. *AI and Ethics*.

Christian Widmer. 2021. Investigation of bias in hate speech classifications. Master's thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.