On the Interaction between Annotation Quality and Classifier Performance in Abusive Language Detection

Holly Lopez Long, Alexandra O'Neil, Sandra Kübler
Indiana University
Bloomington, IN, USA
{hdlopezl,aconeil,skuebler}@iu.edu

Abstract

Abusive language detection has become an important tool for the cultivation of safe online platforms. We investigate the interaction of annotation quality and classifier performance. We use a new, fine-grained annotation scheme that allows us to distinguish between abusive language and colloquial uses of profanity that are not meant to harm. Our results show a tendency of crowd workers to overuse the abusive class, which creates an unrealistic class balance and affects classification accuracy. We also investigate different methods of distinguishing between explicit and implicit abuse and show lexicon-based approaches either over- or under-estimate the proportion of explicit abuse in data sets.

1 Introduction

In recent years, annotation quality has come under closer scrutiny, especially for subjective classification tasks that rely on human judgement. Investigations of unintended bias in abusive language data sets have demonstrated that they are susceptible to sampling and annotation bias (Wiegand et al., 2019; Sap et al., 2019). Although this work provides some guidance for reducing the effects of unintended bias from sampling, it does not provide a clear path forward for mitigating annotation bias. As we need effective ways to curb online hate, we definitely need reliable data sets with high-quality annotations for abusive language detection.

In this paper, we compare annotations by untrained crowd workers with annotations by experts. Our examination demonstrates that labeling differences between crowd workers and experts change the class distribution in the data set and affect classifier performance. We also compare methods for determining explicit and implicit abuse in the data set, and how this affects the interpretation of machine learning experiments. Our paper is structured as follows: Sec. 2 explains our research questions, sec. 3 describes prior work on data sets, annotation procedures and quality, and classification schemes for abusive language, sec. 4 describes our data sets and methodology, sec. 5 discusses our insights into the interaction of annotation quality and classifier performance, and sec. 6 investigates the interaction of explicit and implicit abuse and the interpretation of results. We conclude in sec. 7.

2 Research Questions

We started our investigation by reviewing a random sample of 1 000 posts from the Kaggle competition *Jigsaw Unintended Bias in Toxicity Classification*[1]. Our initial inspection showed that many posts that were considered abusive by crowd annotators (wrt. the final classification used by the competition) were open to interpretation. The examples below show typical "abusive" posts where different interpretations are possible.

1. I do love Yataimura Maru's ramen? It is a perfect food for Portland's long winter. And PDX does kick a little ass.

2. Sorry to have to do this, but just to see if profanity filtering is enabled: fuck.

3. Took this as an opportunity to check back in on The Yard and the floorplans are finally up and they are ATROCIOUS.

The first example is a positive review of a ramen restaurant in Portland that also contains profanity. The second example also uses profanity, but is doing so as a part of a meta-comment about the filters used by the platform. Finally, the last post is a criticism of an apartment complex. Although an insult is used, it is directed toward an object and not

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Proceedings of Recent Advances in Natural Language Processing, pages 868–875
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_099

individuals. These examples show two common tendencies by crowd workers: interpreting profanity as abusive without considering the context and not distinguishing between insults and criticism directed at people and objects. These observations led us to our first research question:

RQ1: How does annotation quality affect characteristics of the data set and subsequently the performance of machine learning approaches? More specifically, we address the following questions:

RQ1.1: How does annotation quality affect the distribution of abusive and non-abusive posts?

RQ1.2: Based on a detailed annotation scheme that distinguishes varieties of non-abuse, how do crowdsourced and expert annotations differ?

RQ1.3: How do annotations by crowd workers and experts influence classification results?

We then turn to the issue of explicit vs. implicit abuse. It is generally accepted that explicit abuse is easier to detect automatically (Wiegand et al., 2019). However, the method that we use to determine whether a post is explicitly or implicitly abusive will result in different splits of the data, and different results of how well a classifier performs on either class. Explicit abuse is often identified via lexicons of abusive expressions. Automatically created lexicons, such as the one by Wiegand et al. (2018), have good coverage, but may overestimate the abusiveness of terms while manually curated lists, such as the one by Razo and Kübler (2020), are more reliable in their selection of abusive terms but may lack coverage. This leads to our second question:

RQ2: How does the method of identifying explicit abuse influence the distribution in the data set and subsequently the interpretation of the classifier's performance?

3 Related Work

Data Sets and Their Development There is an abundance of data sets available for abusive language detection, which represent a variety of approaches for annotating abusive content. While many data sets have relied on large pools of crowd sourced annotators (Zampieri et al., 2019), others have used experts (Waseem and Hovy, 2016). Crowdsourcing annotations is often an attractive option for developing abusive language data sets, since the process often requires a considerable amount of time and labor. The two largest data sets were created for Kaggle competitions: 1) The *Toxic Comment Classification Challenge*[2], which contains 312 737 posts from Wikipedia Talkpages, and 2) the *Jigsaw Unintended Bias in Toxicity Classification* (see Section 4.1) with posts from the platform Civil Comments. Both data sets were annotated by crowd workers. However, using crowd workers can contribute to diminished annotation quality (Hsueh et al., 2009). Waseem (2016) found that amateur annotators were more likely to label a post as hate speech and expert annotations improved machine learning performance. Sap et al. (2019) showed amateur annotators more often labeled African American English posts as abusive, but that priming the amateur annotators for dialect and race reduced annotation bias.

Annotation Schemes One way to maintain annotation quality is to create clear annotation guidelines with rich taxonomies (Vidgen and Derczynski, 2020). Many scholars have developed annotation schema and guidelines that describe different types of abuse in order to better characterize abusive content. Founta et al. (2018) evaluated 7 abuse categories (e.g., offensive, abusive, hateful, aggressive, cyberbullying, spam, and normal), which were then merged into four (e.g., abusive, hateful, spam, and normal) when they found overlap between categories. Zampieri et al. (2019) created a 3-tier scheme in which annotators decided whether a post was abusive, targeted, and whether the target was an individual, a group, or other. Davidson et al. (2017) distinguished hateful content from the casual use of profanity by creating three categories: hateful, offensive (but not hateful), and neither. Current methods overwhelmingly focus on the labeling of abusive posts, often at the expense of accuracy on non-abusive posts.

Investigating Unintended Bias Recent work on abusive language detection has looked at sampling bias in the data. Sampling methods are required to increase the amount of abusive posts in data sets. However, the specific sampling methods used have been shown to create bias. Wiegand et al. (2019) document this bias. Razo and Kübler (2020) build on their work and find that the source of the text (Twitter, Wikepedia, etc.) has more influence on the bias of the data set than the sampling method.

[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

4 Methodology

4.1 Data

We use subsets of the data set from the Kaggle competition *Jigsaw Unintended Bias in Toxicity Classification* with posts from the platform Civil Comments. For the Jigsaw challenge, each comment was annotated by several crowd workers and a mean annotation score of ≥ 0.5 (range: 0.0–1.0) was considered abusive. From this data set, we use two sampling subsets by Razo and Kübler (2020): the first subset of the random boosted sampling sets (a random sample) and the first of the biased topic sampling sets (increasing the number of abusive posts by searching for controversial topics that tend to attract abuse)[3].

4.2 Classifier Settings

For the machine learning experiments below, we follow similar procedures as Razo and Kübler (2020). As Razo and Kübler, we use SVMs, more specifically, the SVC class of Scikit-learn (Pedregosa et al., 2011) with the RBF kernel, the same parameter settings (e.g., C=1000, gamma=0.001), and word 1-3 grams for features. We also perform 5-fold cross validation. Unlike Razo and Kübler, we do not remove punctuation.

5 Investigating Annotation Quality

As discussed in section 2, a cursory inspection of the data sets showed that there was a considerable amount of posts that were annotated as abusive (based on Jigsaw's definition of the challenge), which the expert annotators found questionable. This does not only mean that the classifier learns a model that is disposed towards classifying too many posts as abusive, it also raises the question of whether a more consistent annotation would improve classification results or make the task more difficult to learn (since profanity would have to be disambiguated between abusive and colloquial use). For this reason, we decided to re-annotate the abusive portions of the two data sets. We first present the new annotation scheme in section 5.1, then we describe the resulting changes to the data set in section 5.2 and on the classifier in section 5.3.

5.1 New Annotation Scheme

The new annotation scheme includes 8 categories: explicit, implicit, self-abusive, irony, colloquial, meta, argumentative, and non-abusive. The categories of explicit, implicit, and self-abuse are considered to be abusive, all other categories are non-abusive. These categories were developed using a grounded theory approach (Glaser and Strauss, 1967; Corbin and Strauss, 2014), where the researchers open-coded a small set of instances originally labeled abusive and then consolidated categories and refined definitions. During the open-coding stages, researchers focused on characteristics that non-experts erroneously consider abusive.

We created categories to capture the challenging nuances of language, such as discussions about abuse (meta) and argumentative statements that may be antagonistic to a particular idea or policy, but not abusive toward individuals. While many existing schemes focus on distinctions between different varieties of abuse (Founta et al., 2018), our scheme [4] focuses on non-abusive instances including profanity, etc. The categories with examples are shown in Table 1. The last category is for posts that were originally labeled as non-abusive.

The category *explicit* describes posts that use insults, threats, ethnic/religious slurs, and/or *ad hominem* attacks. This included instances of cyberbullying (e.g., attacking people's appearance/body shape) and other forms of overt hate based on attributes of their identity, such as religion, ethnicity, sexuality, disability, or socioeconomic class.

The category *implicit* is used to indicate that a post degrades individuals or groups of people by alluding to stereotypes or other insulting speech through indirect methods. These posts include the same stereotypes apparent in explicit abuse, but instead of being directly expressed, the abuse is implied in the post.

Self-abuse is used to label posts in which people direct the abuse against themselves. While it could be argued that people should have the right to abuse themselves, we group this category with the other abusive categories because certain types of self-abuse may result in a diminished sense of self-worth (cf. e.g., negative self-talk).

As posts can often belong to several categories, especially since posts are often longer than Twitter posts, we label each post with all applicable labels. However, for all machine learning experiments, we reduced the annotation automatically to a single label per post, either abusive or non-abusive, to keep

[3] Available at https://github.com/danterazo/abusive-language-detection/

[4] https://github.com/hlopezlong/Annotation_Quality/blob/main/AnnotationGuidelines.txt

Category	Example
explicit	*Liberals are just bone stupid. There can be no other rational explanation for their bias and ignorance.*
implicit	*Trump loves his uneducated voters. It sounds like you know a few yourself.*
self-abuse	*I not only missed the point, I missed the headline. I screwed up. I attempted to delete my idiotic comment several times but it keeps reappearing. Stupid is as stupid does and I sure did stupid (to slightly misquote our president).*
irony	Well shit, they drafted a guide. We should all be good now, whew aht a relief...
colloquial	*DARPA, the subdivision of the Defense Department in charge of devising Really Scary Shit That's Never Been Seen on Earth Before, aka the inventors of the internet.*
meta	*The slurs against Hillary should be stopped—- it's time to confront them at every appearance. We all have seen that to ignore them as too ridiculous isn't effective, i.e. Saddam had WMDs, Saddam caused 9/11, Obama is a Muslim, etc.*
argumentative	*Great story. Franke tried to expose corruption and ends up murdered. Problematic interrogation tactics by OSP. Can't wait for more info on this case and final proof of the real murderer, if this man is not responsible. Reinterview Franke's brother. He used to comment on WW now and then.*
non-abusive	*Perhaps they're not legitimate, civil comments.*

Table 1: Annotation categories and examples.

consistency with prior experiments. If a post contains any of the abusive categories, it is considered abusive. All posts that contain only non-abusive categories (i.e., irony, colloquial, meta, argumentative, non-abusive) are considered non-abusive. For the question on explicit vs. implicit abuse, we only examine instances considered explicit and/or implicit abuse but ignore the other categories.

5.2 Effect on Annotations

We first look at the effects of re-annotating the *abusive* posts from the original annotations, since we noticed previously a large number of false positives in the annotations. However, note that the annotation scheme can and should be applied to all posts. The two data sets were re-annotated by the first two authors, with each author being responsible for one data set. In order to ensure consistency, both annotators collected all posts that raised questions; these posts were discussed by all authors, and a consensus was reached.

When we compare the original annotations (by non-experts) with our expert annotations, we see the following trends: Although the overall agreement between expert and non-expert annotations remains high across both samples (95.3%), agreement is significantly lower (45.6%) when looking only at re-annotated instances. For the random boosted sampling set, labels between experts and non-experts only have an agreement of 46.9%. On the biased topic sampling data set, expert and non-expert annotations agree 44.7% of the time[5].

Category	Count
explicit	1 172
implicit	402
self-abuse	7
total abusive:	1 581
argumentative	1 514
colloquial	40
irony	42
meta	273
non-abusive	6
total non-abusive:	1 875

Table 2: Updated counts of orig. abusive posts.

When looking at the distribution of labels in the re-annotated posts shown in Table 2, the large majority of disagreement between annotators are posts that the expert annotators consider argumentative instead of abusive, i.e., posts expressing criticism or disagreement, without targeting insults or criticism at individuals or groups of individuals.

5.3 Effect on Evaluation Results

Since the re-annotation has a significant impact on the annotations, and especially on the skewing between the abusive and non-abusive class, we expect

[5]Since our annotation scheme differs from the the crowd workers', we could not compute inter-annotator agreement.

Sample	Annotation	Category	% in set	Precision	Recall	macro-F
topic	expert	non-abusive	95.90	96.11	99.80	97.92
		abusive	4.10	54.22	5.49	9.97
		overall	100.00	75.16	52.64	53.94
	crowdsourced	non-abusive	90.83	92.34	99.20	95.65
		abusive	9.17	70.04	18.48	29.25
		overall	100.00	81.19	58.84	62.45
random	expert	non-abusive	96.20	96.63	99.45	98.02
		abusive	3.81	47.00	12.35	19.56
		overall	100.00	71.82	55.90	58.79
	crowdsourced	non-abusive	91.89	93.55	98.96	96.18
		abusive	8.11	65.83	22.69	33.75
		overall	100.00	79.69	60.82	64.96

Table 3: Precision, recall, and macro-averaged F for non-abusive and abusive posts for the retrained classifier.

that they will also have a considerable effect on the difficulty of the task and consequently the classification quality. We investigate the general question of how exactly the re-annotation affects classification, and we focus on two specific questions: 1) How does the re-annotation affect the results of a classifier trained on the new gold standard? And 2) How does the new gold standard affect the evaluation of classifications trained on the original data from Razo and Kübler (2020)? In other words, is the classifier potentially more consistent than the crowd workers?

5.3.1 Evaluating a Retrained Classifier

To determine how the new annotation scheme affects classification accuracy, we train and test the SVM using the same parameter settings as Razo and Kübler (2020) (see section 4.2). We also use two of their data sets, but with the re-annotations of the original abusive posts.

The overall results show that the crowdsourced annotations are easier to learn; they result in higher scores across all evaluation measures than their expert annotation counterparts, regardless of sampling methods. For topic biased sampling, the macro-averaged F-score decreases from 62.45 for the crowdsourced annotations to 53.94 for the expert annotations. For random boosted sampling, the decrease is comparable, from 64.96 to 58.79. One of the reasons can be found in the class skewing: for both samples, the percentage of abusive posts in the sample decreases by about 5%. Thus, the skewing is even more extreme in the re-annotated data. However, the decrease in the classifier's F-score is about twice as much, which leads us to the assumption that the simpler cases were moved from the abusive class to the non-abusive one. This may also have an effect on the distinction between explicit and implicit abuse, see section 6.

When looking more closely at the evaluation measures for abusive and non-abusive posts in Table 3, we observe that the re-annotation of the abusive posts leads to decreased precision and recall for abusive posts. For the topic biased sample, precision decreases from 70.04% to 54.22%; for the random boosted sampling, the decrease is from 65.83% to 47.00%. Recall is affected even more dramatically, it drops from 18.48% to 5.49% for topic biased sampling, and from 22.69% to 12.35% for random boosted sampling. However, at the same time, the re-annotation leads to an improvement of those same measures for non-abusive posts and more specifically to a considerable improvement of precision: For biased topic sampling, precision increases from 92.34% to 96.11% and for random boosted sampling from 93.55% to 96.63%. These changes are unsurprising given the changes in class skewing. Additionally, and more importantly, the task of identifying abusive posts has become more difficult. Of the 314 instances (across both data sets) where the classifier agrees with the crowdsourced annotation rather than the expert one, 78.37% are argumentative and not directed at people, 19.44% are meta comments about abuse, and 0.63% are colloquial use of profanity. Once these posts are labeled as non-abusive, the classifier basically needs to disambiguate between meta comments like, "...I have voted "not civil" on posts i deeply agree with but which call the other person "idiot" or some such." and abusive comments such

		Crowd		Expert	
		Prec.	Rec.	Prec.	Rec.
topic	non-ab.	0.00	0.00	57.26	84.42
	ab.	100	18.48	53.39	22.07
rand.	non-ab.	0.00	0.00	55.82	81.30
	ab.	100	22.69	56.25	27.20

Table 4: Evaluating results by Razo and Kübler (2020) against both gold standards (orig. abusive posts only).

as, "The women is an IDIOT and if left in office she will destroy German identity . . . ".

5.3.2 Re-Evaluating Prior Results

To better understand the impact of annotation quality, we re-evaluate the classification results by Razo and Kübler (2020) on the two data sets, i.e., we contrast the two gold standards in evaluation. This means that we use the predictions of the classifier that was trained on the crowdsourced training data, and compare this to the new gold standard created by expert annotations.

The re-evaluation is performed on the subset of the original posts only (since those are re-annotated). Thus, precision for both categories and recall for non-abusive are meaningless, either 0.00 or 100.00 on the crowdsourced annotations. The results of the re-evaluation in Table 4 show that recall on the abusive class increases when evaluated against the expert annotations, from 18.48% to 22.07% for topic biased sampling and from 22.69% to 27.20% for random boosted sampling. This means that more of the posts that the classifier annotated as abusive are abusive based on the experts opinion. Thus, partly, the classifier is sensitive to distinctions that the crowd workers may have neglected. However, a look at precision of around 55% for both classes and both samples shows that the classifier creates many false positives and is still far from having learned the more conservative expert regularities.

6 Investigating Explicit vs. Implicit Abuse

Now we turn to the distinction between explicit and implicit abuse. It is generally accepted that explicit abuse is easier to detect than implicit abuse. However, making this distinction is not a simple task. In general, lexicons of abusive words are used to determine the explicitly abusive posts; a post is considered explicit abuse if one of the lexicon words occurs in the post. Wiegand et al. (2018) describe a method for automatically extending a base lexicon into a larger lexicon of abusive words. Their base list contains 551 words, their extended list 2 989 words. Razo and Kübler (2020) show that both the base and the extended lexicon cover a large proportion of posts that were labeled non-abusive by crowd workers. They manually checked the base lexicon and reduced it to 151 words.

Since we now have expert annotations for the abusive posts, we can investigate how the distribution of explicit and implicit posts in the two gold standards differs from those based on the three lexicons. We compare the proportions of posts from each gold standard labeled as explicit and implicit abuse with the lexicon approaches used by Razo and Kübler (2020) and Wiegand et al. (2018). Table 5 shows the proportions in each data set.

Distributions of implicit and explicit abuse in Table 5 show that within the abusive category as defined by experts, the methods for determining explicit vs. implicit abuse result in very different distributions. The extended lexicon by Wiegand et al. (2018) results in the highest proportion of 92.80% (topic) and 91.55% (random) explicitly abusive posts. Our expert annotation and the Wiegand base lexicon result in similar proportions between 67.15% (random) and 79.74% (topic) of explicit abuse while the manually checked lexicon only groups 46.95% (topic) / 34.82% (random) of the posts as explicit. These lower numbers are to be expected since the authors state that the manual lexicon is very small and thus has coverage issues.

However, the similarity in proportions raises the question whether the Wiegand base lexicon and the manual annotations choose the same posts, or just the same proportion of posts. We checked the overlap of posts that were labeled explicit or implicit by both (not shown in table). For the topic sample, 75.96% of explicit posts annotated by experts can be found in the posts extracted using the base lexicon. The random sample shares a smaller proportion of explicit posts (66.94%) than the topic sample. There is also a smaller proportion of overlap between annotations and the base lexicon among implicit posts. 24.90% of posts from the topic sample and 32.89% of posts from the random sample can be found in the implicit posts using the base lexicon method. This shows very clearly how different the samples of explicit and implicit abuse are based on the different methods.

Sample	Gold standard	Lexicon	In abusive		In all	
			Explicit	Implicit	Explicit	Implicit
Topic	expert	annotations	79.74	20.26	2.85	1.25
		Razo manual	46.95	55.05	22.50	77.50
		Wiegand base	75.73	24.27	58.15	41.85
		Wiegand extended	92.80	7.20	89.39	10.61
	crowds.	Razo manual	39.15	60.85	22.50	77.50
		Wiegand base	74.65	25.35	58.15	41.85
		Wiegand extended	98.84	6.16	89.39	10.61
Random	expert	annotations	70.00	30.40	3.01	0.77
		Razo manual	34.82	65.18	16.68	83.32
		Wiegand base	67.15	32.85	45.98	54.02
		Wiegand extended	91.55	8.67	79.64	20.36
	crowds.	Razo manual	31.69	68.31	16.68	83.32
		Wiegand base	68.25	31.75	45.98	54.02
		Wiegand extended	91.55	8.45	79.64	20.36

Table 5: Distribution of implicit and explicit posts across lexicon methods and annotations of abusive posts.

Sample	Lex.	Cat.	Rec.	F
topic	expert	explicit	6.14	11.57
		implicit	4.02	7.72
	base	explicit	6.44	12.10
		implicit	2.51	4.90
random	expert	explicit	13.95	24.49
		implicit	5.88	11.11
	base	explicit	13.50	23.79
		implicit	10.00	18.18

Table 6: Effect of definitions of implicit and explicit categories on performance on the *abusive class*.

We then investigate how these different decisions affect classification results. For this, we use the same classification results from section 5.3, and we evaluate the subsets against our expert annotations. The subsets consist of only explicitly or implicitly abusive posts, based on either expert annotations or the base lexicon. The results are shown in Table 6. Since both precision and recall for the non-abusive class are 0.00 (and precision for the abusive class 100.00), we only report recall for the abusive class. A comparison of the recall results shows that there are differences between the expert annotation and the lexicon approach. However, for the two samples, they go in two different directions: For the topic biased sample and the explicit category, the classifier performs better based on the lexicon subset while for the random boosted sample, it performs better based on the expert annotations. The trends for implicit abuse also show this difference, but in the opposite direction. Part of this discrepancy is certainly due to the low overlap in the explicit/implicit subsets in the random sample. It is also clear that the definition of these categories has a significant influence on the interpretation, given the sizable differences in recall, thus requiring future work in this area.

7 Conclusion and Future Work

Our investigation illustrates the effect of diminished annotation quality on machine learning performance. Crowd workers and expert annotators disagreed on approximately a third of the posts originally labeled abusive. Disagreement often occurred with profanity and when targets were not individuals. The method for identifying implicit and explicit abuse leads to significant discrepancies between the explicit and implicit classes and affects evaluation.

Our work shows the need to improve annotation quality. This is only the tip of the iceberg, though. Verbal abuse can only be identified within a cultural context, but there exist so many different subcultures that any annotator, independent of their being sensitized to the nuances of abuse, may not be able to identify abuse if they are not part of that subculture. We will investigate using annotators with different backgrounds along with methods to distinguish between disagreement based on inattention or lack of sensitization from disagreement based on cultural backgrounds.

References

Juliet Corbin and Anselm Strauss. 2014. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, Montréal, Canada.

Antigoni 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. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 491–500, Palo Alto, CA.

Barney G Glaser and Anselm L Strauss. 1967. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Sociology Press.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, CO.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the 4th Workshop on Online Abuse and Harms (WOAH)*, pages 70–78, Online.

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.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12).

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, TX.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, CA.

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*, pages 602–608, Minneapolis, MN.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1046–1056, New Orleans, LA.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420, Minneapolis, MN.