Investigating Sampling Bias in Abusive Language Detection

Dante Razo, Sandra Kübler
Indiana University
{drazo,skuebler}@indiana.edu

Abstract

Abusive language detection is becoming increasingly important, but we still understand little about the biases in our datasets for abusive language detection, and how these biases affect the quality of abusive language detection. In the work reported here, we reproduce the investigation of Wiegand et al. (2019) to determine differences between different sampling strategies. They compared boosted random sampling, where abusive posts are upsampled, and biased topic sampling, which focuses on topics that are known to cause abusive language. Instead of comparing individual datasets created using these sampling strategies, we use the sampling strategies on a single, large dataset, thus eliminating the textual source of the dataset as a potential confounding factor. We show that differences in the textual source can have more effect than the chosen sampling strategy.

1 Introduction

Abusive language detection has become an important problem, especially in a world where #BlackLivesMatter, and where abusive posts on social media need to be found and deleted automatically. However, we also know that the datasets that we currently use for training classifiers are all biased in some way or another. Wiegand et al. (2019) present one of the first investigations into into bias in different datasets for abusive language detection for English. They compare characteristics of 6 datasets, based on their underlying sampling strategy, their proportion of abusive posts, and the proportion of explicit abuse. The proportion of abusive posts is important for classifiers: If that proportion is too small, classifiers tend default to the majority class baseline. Thus, creators of datasets use a range of strategies to increase the number of abusive posts in their data. The specific strategy can have an influence on the proportion of explicitly abusive posts, which tend to be easier to identify as abusive, and overall on classifier performance. We are interested in understanding this interaction better.

Wiegand et al. (2019) distinguish between boosted random sampling and biased topic sampling. Boosted random sampling is based on a complete sample, for example all tweets of a specific time frame. Then, the number of abusive posts is boosted using different methods, for example by adding more posts by users who have been blocked for being abusive. Biased topic sampling, in contrast, samples posts from specific topics, such as soccer or Islam, which are known to cause a considerable amount of abuse.

Wiegand et al. (2019) argue that the type of sampling strategy introduces bias into the dataset, and we can assume that the two sampling strategies create different biases: Random boosted sampling may create a bias towards specific authors but with a widespread range of topics, and biased topic sampling may create a bias towards specific topics, and potentially specific authors. However, we are often unaware of the exact biases present in such datasets. This is important because first results on debiasing datasets show that these methods work best when we know which bias is present (He et al., 2019).

In our work, we focus on reproducing the results by Wiegand et al. (2019) and providing a closer look at the different sampling strategies. While Wiegand et al. (2019) normalize performance by using a single classifier on all datasets, they do not normalize across different text types. Thus, the two sampling strategies have been used on different datasets, which leaves open the question to what degree the differences in bias are due to textual characteristics (Wikipedia talkpages, Twitter feed, Facebook posts), or to the sampling strategies. Consequently, we repeat their experiments applying both sampling techniques to the same dataset.

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 70–78

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

We use two datasets from Kaggle competitions, of sufficient size to enable us to sample from the *same dataset* and obtain smaller subsets based on different sampling strategies. We also add an investigation into two variants of biased topic sampling and the out-of-vocabulary rate of the resulting subsets.

The remainder of the paper is structured as follows: Section 2 explains our research questions, section 3 provides an overview of related work on bias in abusive language detection data, and section 4 discusses our experimental setup, including datasets, lexicons, sampling strategies, the classifier, and evaluation. In section 5, we discuss our findings, and in section 6, we conclude.

2 Research Questions

When reproducing the investigation by Wiegand et al. (2019), we focus on the following questions:

1. Does repeated sampling from a dataset change characteristics of the data?

   We first need to investigate how diverse the Kaggle datasets are, i.e., to what extent sampling a fairly small subset will change the distribution and difficulty of the dataset. Thus, we create 3 sampled subsets and compare their results.

2. Are there performance differences between boosted random sampling and biased, topic-based sampling?

   This is a replication of the question by Wiegand et al. (2019), but we first compare the two sampling strategies on samples from the *same* underlying dataset, the original Kaggle dataset also used in their experiments (see section 4.1 for details on the datasets), which is originally based on boosted random sampling.

   Additionally, we repeat the experiment on another, larger Kaggle dataset for abusive language detection.

3. How dependent are results on the topic used for sampling?

   Since the original Kaggle dataset is based on Wikipedia talkpages and thus covers topics different from the one covered in other datasets, we could not use the list of topics used by previous approaches for biased topic sampling (Kumar et al., 2018; Waseem and Hovy, 2016; Warner and Hirschberg, 2012). This leads to the question how dependent results are on the choice of topics. We compare the wide range of topics we used for the previous question to a setting where we use only one specific term to sample.

4. To what degree does the proportion of explicit abuse and the OOV rate correlate with performance?

   Wiegand et al. (2019) also ranked datasets based on the proportion of explicitly and implicitly abusive language. We have a closer look at this distinction, along with looking at the OOV rate of instances.

3 Related Work

Wiegand et al. (2019) were among the first to investigate bias in datasets used for abusive language detection. They compared 6 different datasets and found topic and author bias, which was introduced by the sampling method used to create the datasets. As a method to avoid biased evaluation, they recommend cross-domain classification, i.e., using different datasets to train and test an approach.

Additionally, van Rosendaal et al. (2020) investigate methods for boosting abusive language when creating datasets while at the same time maintaining a good spread of topics. They suggest concentrating on controversies and describe two specific methods: For Twitter data, they suggest using the most frequent hashtags over a time period. And for Reddit, they suggest using posts that have a similar number of up- and down-votes, a sign for the controversial nature of these posts.

Park et al. (2018) discuss methods to decrease the gender bias in abusive language detection. They suggest 3 methods for debiasing, which successfully reduce gender bias in their experiments: debiasing word embeddings, gender swap data augmentation, and fine-tuning using a larger corpus.

Sap et al. (2019), in contrast, focus on racial bias, which is originally introduced by annotator's insensitivities to African-American English (AAE), but is then propagated via a trained classifier learning this bias. Sap et al. (2019) show that priming the annotators for dialect and race of the tweet's producer results in fewer AAE posts being labeled abusive. Davidson et al. (2019) provide a more in-depth analysis, showing that the bias also holds when comparing tweets containing the keywords "n*gga" and "b*tch".

There are also approaches to eliminate bias from datasets. For example, Badjatiya et al. (2019) present a method to identify and replace bias sensitive words.

4 Experimental Setup

4.1 Datasets

We use the largest datatset from the sets used by Wiegand et al. (2019), the dataset from the Kaggle *Toxic Comment Classification Challenge*[1]. This dataset is an extension of the dataset by Wulczyn et al. (2017). The dataset contains 312 737 posts from Wikipedia Talkpages. It was created using random boosted sampling; the authors boosted the number of abusive posts by sampling posts from "users who where blocked for violating Wikipedia's policy on personal attack" (Wulczyn et al., 2017). We consider all posts abusive which are marked as either "toxic" or "severely toxic", following Wiegand et al. (2019). We will refer to this dataset as the *original Kaggle* set.

Additonally, we use the dataset from the Kaggle competition *Jigsaw Unintended Bias in Toxicity Classification*[2] with posts from the platform Civil Comments. The dataset contains 1 804 874 posts. Following Jigsaw's documentation, we consider every post with a target value of ≥ 0.5 abusive. We chose this dataset mainly because of its size since it gives enough posts for the sampling process, but also because the data are from a different domain than the first data set. We will refer to this dataset as the *large Kaggle* set.

4.1.1 Data Preprocessing and Features

For both datasets, we only use the posts and the abusive rating. We used the Scikit-learn (Pedregosa et al., 2011) tokenizer to tokenize the posts and then removed punctuation.

We use 5-fold cross-validation on all datasets, and we use word 1-3-grams as features.

4.2 Lexicons

Following Wiegand et al. (2019), we use a lexicon-based approach to determine whether a post is explicitly or implicitly abusive. As lexicons, we consider the base and extended lexicon by Wiegand et al. (2018). The base lexicon was created from negative polar expressions and annotated for abusive terms via crowdsourcing. This lexicon was used in a classifier to create the extended lexicon.

However, a manual inspection showed that many of the words in the base lexicon were not offensive. For this reason, we created a manually-vetted version of this lexicon[3]. Three native speakers were asked to rate each word in the base lexicon as either non-abusive, mildly abusive, or definitely abusive. For our manually-vetted lexicon, we consider all words abusive that 2 or three of our annotators have considered mildly or highly abusive. The base lexicon contains 551 abusive entries, the extended lexicon has 2 989 entries, and our manually-vetted lexicon 151 abusive words.

The native speakers disagreed with the original classification of 269 words from the base lexicon. Examples of words deemed inoffensive include "aloof", "chonky", and "gossip". There were only 6 words that were highly abusive, according to all three judges. Among them are n*gger, f*g, and c*nt.

4.3 Generating Sampling Variants

Our experiments utilize three types of sampling: boosted random sampling, biased topic sampling, and biased topic sampling with a narrowly defined topic (see below). For each sampling type, we sampled three subsets of 20 000 posts per dataset.

Random Boosted Sampling Since the original Kaggle dataset is based on random boosted sampling, we can use basic random sampling from the dataset to obtain our boosted random samples. For the large Kaggle set, it is unclear how these posts were collected, but it is more likely to be a variant of random boosted sampling than biased topic sampling, thus we used the same strategy as for the original Kaggle set.

Biased Topic Sampling Since the original Kaggle dataset is extracted from Wikipedia talkpages, the topics covered in the dataset are different from the topics in previous datasets using biased, topic-based sampling (Kumar et al., 2018; Waseem and Hovy, 2016; Warner and Hirschberg, 2012). Consequently, we had to create our own list of topic words. We created a list of (non-abusive) topic words covering a wide range of topics found in

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

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

[3] The manually-vetted lexicon is available at https://github.com/danterazo/abusive-language-detection/blob/master/data/lexicon.manual.csv

Set	Category	% in set	Precision	Recall	F1
set 1	not abusive	90.51	95.39	99.05	97.19
	abusive	9.49	85.73	54.50	66.56
set 2	not abusive	90.62	95.35	99.02	97.16
	abusive	9.38	85.00	53.44	65.62
set 3	not abusive	90.50	95.17	99.03	97.06
	abusive	9.50	84.92	52.13	64.60

Table 1: Results of repeating the random subset sampling process from the original Kaggle dataset.

Set	Category	% in set	Precision	Recall	F1
set 1	not abusive	91.89	93.55	98.96	96.18
	abusive	8.11	65.83	22.69	33.75
set 2	not abusive	91.83	93.41	98.86	96.06
	abusive	8.17	62.75	21.54	32.07
set 3	not abusive	91.94	93.66	99.14	96.33
	abusive	8.06	70.58	23.51	35.27

Table 2: Results of repeating the random subset sampling process from the large Kaggle dataset.

the Kaggle datasets, and which are known to incite abuse. The topics include, but are not limited to, politics, religion, and social justice initiatives, example words are "immigration", "muslim", and "feminism"[4].

Narrow Topic We chose the name "Trump" as our topic for the biased topic sampling with a narrow basis, assuming that the discussions around the last presidential elections will have incited abusive comments (the large Kaggle dataset covers posts from 2015 through 2017).

4.4 Classifier

We deviated from Wiegand et al. (2019) and used Support Vector Machines as our classifier. We used the SVC implementation of Scikit-learn (Pedregosa et al., 2011).

Scikit-learn's GridSearchCV was used to optimize our model parameters in an initial experiment. We then used the following optimal parameters for all consecutive experiments: regularization parameter: 1000, gamma: 0.001, and the radial basis function (RBF) kernel.

4.5 Evaluation

We report accuracy, macro-precision, macro-recall, and macro-F1 scores. For all classification experiments, we report averages over 3 samples and 5-fold cross-validation on each sample. For all statistics, we report averages over the 5 folds of the first sample.

[4]The list of words is available at https://github.com/danterazo/abusive-language-detection/blob/master/data/wordbank.py.

5 Results

5.1 Repeated Subset Sampling

Here we investigate the consistency of datasets sampled from the two Kaggle datasets. We create 3 randomly sampled datasets of 20 000 posts, and then perform 5-fold CV on each set. Note that this type of sampling is different from the sampling investigated in the next sections; here the goal is to reduce the size of the dataset to a uniform, small size, and we need to determine how much variation we should expect from this random subset sampling.

The results for the original Kaggle set are shown in Table 1 and for the large Kaggle set in Table 2. They show that all three samples have a very similar distribution of classes. Both datasets also show a similar performance of the classifier on the majority class. For the minority class, in contrast, there are differences in the range of 2% absolute: The original dataset shows a decrease in the F-score from 66.56 in set 1 to 65.62 in set 2 and 64.60 in set 3. For the large Kaggle set, set 3 shows a noticeably higher performance (35.27) than set 1 and 2 result, which are similar in F-scores (33.75 and 32.07).

Given these results, we decided to use repeated subset sampling for all the remaining experiments,

	Boosted random sampling				Biased topic sampling			
Category	% in set	Precision	Recall	F1	% in set	Precision	Recall	F1
Not Abusive	90.54	95.31	99.03	97.13	93.85	96.14	99.23	97.66
Abusive	9.46	85.21	53.32	65.59	6.15	76.91	39.23	51.95
Accuracy		90.26				86.53		

Table 3: Comparing boosted random sampling and biased topic sampling on the original Kaggle dataset.

	Boosted random sampling				Biased topic sampling			
Category	% in set	Precision	Recall	F1	% in set	Precision	Recall	F1
Not Abusive	91.89	93.54	98.99	96.19	90.72	92.23	99.17	95.57
Abusive	8.11	66.38	22.58	33.70	9.28	69.12	18.23	28.85
Accuracy		92.79				91.67		

Table 4: Comparing boosted random sampling and biased topic sampling on the large Kaggle dataset.

thus all classification results below are based on an average over 3 subsets.

5.2 Comparing Random Boosted Sampling and Biased Topic Sampling

This question reproduces the comparison of the two sampling strategies by Wiegand et al. (2019), boosted random sampling and biased sampling (see section 4.3).

The results of this set of experiments for the original Kaggle set are shown in Table 3. A first look at the proportion of abusive posts in the datasets shows that the boosted random sampling results in 9.46% abusive posts while the biased topic sample reaches a lower percentage of 6.15%. This is directly reflected in accuracy, which is lower for the biased topic sample by about the same margin. However, a look at the large Kaggle set in Table 4 shows that the lower rate of abusive posts is not due to the biased topic sampling: In the large Kaggle set, the biased topic sample shows a higher rate of abusive posts than the boosted random set (9.28% vs. 8.11%).

It is also obvious that in both Kaggle sets, the abusive class in the biased topic sample is considerably harder to detect than in the boosted random sample: In the original Kaggle set, the biased topic sample reaches an F-score of 51.95 vs. 65.59 for the boosted random sample. For the large Kaggle set, the biased topic sample reaches 28.85 vs. 33.70. This trend is independent of the distribution of abusive and non-abusive posts, and it mirrors the findings of Wiegand et al. (2019) that the biased topic sampled datasets reach lower F-scores.

However, the differences that we have found are mostly distinct from those found by Wiegand et al. (2019), shown in Table 5. While Wiegand et al. (2019) found that biased topic sampling tends to lead to higher proportions of abusive posts, our samples show that the difference is minimal, thus pointing to the hypothesis that the data source has more influence on the proportion of abusive language than the sampling strategy. They also found that boosted random sampling leads to higher F-scores (with the exception of the Waseem set, whose high F-score they trace back to the topic and author biases in this dataset). The same trend can be found in our samples, but to a much smaller degree: In our samples, the difference is about 1%, the most extreme difference in the datasets by Wiegand et al. (2019) is around 18%, when comparing the Kaggle and Kumar datasets (see Table 5, copied from their paper). This again points to the data source as the main determinant of classifier performance.

5.3 Comparing Wide and Narrow Topic Definitions for Biased Topic Sampling

Given the high performance we obtained on the biased topic sampling on the large Kaggle set, we decided to investigate this point more deeply. We are interested in how the definition of the topic, and more specifically the scope of the topic affects the distribution of abusive and non-abusive posts as well as the performance on this dataset. The first set of experiments for biased topic sampling uses a widely defined set of topics, including politics, religion, and social justice initiatives. Consequently, we created narrow topic samples by focusing on posts that mention the name “Trump”. Note that this experiment is only possible on the large Kaggle set since the original Kaggle set is too diverse in topics and too limited in size to support the sam-

dataset	source	sampling	# posts	% abusive	F1	% explicit
Kaggle	Wikipedia	random	312 737	9.6	88.2	76.9
Founta	Twitter	random	59 357	14.1	87.3	75.9
Razavi	diverse	random	1 525	31.9	83.3	64.7
Warner	diverse	biased	3 438	14.3	71.8	51.3
Waseem	Twitter	biased	16 165	35.3	80.5	44.4
Kumar	Facebook	biased	15 000	58.1	70.4	32.7

Table 5: Dataset characteristics, from (Wiegand et al., 2019, p. 604).

	Wide topic sampling				Narrow topic sampling			
Category	% in set	Precision	Recall	F1	% in set	Precision	Recall	F1
Not Abusive	90.72	92.23	99.17	95.57	86.38	89.50	98.21	93.65
Abusive	9.28	69.12	18.23	28.85	13.62	70.41	26.91	38.93
Accuracy		91.67				88.50		

Table 6: Comparing topic sampling with wide or narrow topic scope on the large Kaggle set.

pling of a narrow topic.

The results for this experiment are shown in Table 6. We repeat the results for biased topic sampling using the wider range of topics from Table 4 for ease of comparison. Not unexpectedly, given the definition of the narrow topic, our sets using narrow topic sampling have a higher proportion of abusive posts compared to the wider topic sampling (13.62% vs. 9.28%). This means that the proportion of abusive posts is closer to the trend that Wiegand et al. (2019) observed, but well below two of the three datasets using biased topic sampling (Waseem with 35.3% and Kumar with 58.1%).

In terms of classifier performance, narrow topic sampling yields higher precision (70.41% vs. 69.12%) and recall (26.91% vs. 18.23%) for abusive posts. There are two possible explanations: Either the higher percentage of abusive posts in training boost performance on this class, or the abusive posts in this set is more consistent in that those posts lean towards explicit abuse. Since the F-score on the non-abusive class is lower for the narrow topic samples (93.65 vs. 95.57%), it is less likely that this sample is more homogeneous. We will investigate the latter aspect below.

The results for the narrow topic are closer to the trend reported by Wiegand et al. (2019) that biased topic sampling reaches lower F-scores. This shows that the definition of the topics included for sampling also have a considerable effect on results.

5.4 Explicit and Implicit Abuse

Wiegand et al. (2019) have also looked at the proportion of explicit vs. implicit abuse on the abusive portion of the datasets, reported in the final column in Table 5. Explicit abuse means that the abusive post contains abusive words; the abuse in implicitly abusive posts is conveyed, for example, "via negation, sarcasm, or negative stereotypes" (Wiegand et al., 2019). They determine explicit abuse using their automatically created lexicon (Wiegand et al., 2018)[5]. We use the two versions of the lexicon by Wiegand et al. (2018) and ours discussed in section 4.2.

In Table 7, we show the proportions of abusive posts in the first sample per sampling condition, based on all 3 lexicons and the original Kaggle set. The first column corresponds to the proportion of explicit abuse in the abusive posts only. The second column shows the proportion in non-abusive posts, and the third column shows the proportion of abuse in *all* posts of the sample. We added the second and third column after a first look at the proportion of abusive posts when using the Wiegand extended lexicon: In all three sampling conditions, the proportion of explicit abuse based on this lexicon is 90.78% or higher. This hints at a significant amount of non-abusive words being included in the lexicon; i.e., over-generation. We test this by looking at the proportion of abusive words in the non-abusive posts and in all data of a sample. If there is a significant proportion beyond the proportion of abusive posts, we have an objective corroboration of our assumption, independent of human judgment. In this setting, using the Wiegand extended lexicon,

[5]More specifically, they use the extended version (p.c. M. Wiegand).

	Abusive		Non-abusive		All	
Lexicon	Random	Wide Topic	Random	Wide Topic	Random	Wide Topic
Wiegand extended	90.78	95.16	76.86	90.83	78.18	91.09
Wiegand base	79.83	84.09	41.32	59.83	44.98	61.30
manual	64.03	65.22	14.76	25.32	19.43	27.75

Table 7: Proportion of explicit abuse in different samples of the original Kaggle dataset.

	Abusive			Non-abusive			All		
Lexicon	Rand.	WT	NT	Rand.	WT	NT	Rand.	WT	NT
Wiegand extended	91.55	93.84	95.07	78.59	88.95	87.37	79.64	89.34	88.30
Wiegand base	68.25	74.65	76.28	44.01	56.48	56.77	45.98	58.15	59.42
manual	31.69	39.15	39.39	15.36	20.82	19.94	16.68	22.50	22.58

Table 8: Proportion of explicit abuse in different samples of the large Kaggle dataset (WT = wide topic, NT = narrow topic).

the proportion of explicit abuse is 78.18% in all posts and 76.86% in non-abusive posts, thus corroborating our assumption.

Overall, we see that the three versions of the lexicon have a significant influence on the proportions of abuse: If we use the Wiegand base lexicon, the proportion of explicit abuse in the whole sample ranges between 44.98% and 61.30%, which is still high given that the proportion of abusive posts in these samples are 9.50% for boosted random sampling and 6.10% for biased topic sampling. For the manually pruned lexicon, the proportions range between 14.76% and 27.75%. This shows the importance of having a high-quality lexicon rather than a large scale list.

When we focus on the manual lexicon and the two sampling methods, we see that the proportion of explicit abuse in both types of samples is very similar, 64.03% for the random sample and 65.22% for the biased topic sample. Thus, the proportion of explicit abuse cannot be the reason for the performance differences we have seen across sampling types for the abusive class.

Table 8 shows the proportions for the large Kaggle set. While we see similar trends with regard to the choice of lexicon, we also see a considerable difference between boosted random sampling and the two biased topic sampling strategies: Based on the manual lexicon, random sampling results in 31.69% explicitly abusive posts while the biased topic sampling strategies reach 39.15% and 39.39%. This may be an explanation of the difference in F-scores on the abusive class in Table 4, but it does not explain the difference in F-scores of around 10 points between the two biased topic sampling methods in Table 6.

When comparing the proportions of explicit abuse across the two Kaggle sets, we see different trends, with more similarities across sampling strategies in the original Kaggle set, and major differences in the large Kaggle set. This indicates that the proportion of explicit and implicit abuse can be more dependent on the text sources than on the different sampling strategies.

5.5 Out-of-Vocabulary Rates

We calculated out-of-vocabulary percentages for both datasets on three conditions: all data, abusive posts only, and non-abusive posts only. These percentages are shown in Table 9. In a way, the OOV rates give us an indication of the topic diversity in given sample. It is to be expected that the original Kaggle set has a higher proportion of OOV words than the large Kaggle set since Wikipedia talkpages cover a wider range of topics than the comments in the large Kaggle set. However, it is interesting to see that this is also true for the biased topic sampling using the wide definition of topics. Here, one would expect the difference between the datasets to be smaller since we choose posts using the same list of topics. However, it is possible that the larger size of the large Kaggle set simply provides more variety to choose from.

It is surprising that the narrow topic has a higher proportion of OOV than the wide topic. In general, abusive posts have the highest OOV rate, independent of sampling strategies. These trends indicate that abusive language seems to be more creative in word choice, as are posts concerning Trump.

In terms of predictive power with regard to clas-

Sampling	original Kaggle			large Kaggle		
	all	non-abusive	abusive	all	non-abusive	abusive
Random	4.24	4.18	12.24	2.66	2.77	9.26
Wide Topic	3.17	3.19	9.65	1.88	2.02	5.27
Narrow Topic				2.17	2.27	7.36

Table 9: Out-of-vocabulary statistics for the two Kaggle datasets.

sifier performance, the OOV rate is also not useful. For example, the narrow topic sampling in the large Kaggle set results in a higher F-score by 10% absolute in comparison to the wide topic samples. However, in terms of OOV rate, the more difficult dataset has a lower OOV rate. Thus, the best predictor of classifier performance is the proportion of abusive posts in a sample. But the differences across datasets are generally larger than the differences across sampling conditions, again stressing that the textual sources of the datasets have more influence on classifier performance than sampling strategies.

6 Conclusion and Future Work

We have investigated the interaction between different sampling strategies with classification results for abusive language detection datasets. We have reproduced the two sampling strategies distinguished by Wiegand et al. (2019), boosted random sampling and biased topic sampling, but we applied them to the same dataset, in order to eliminate the differences resulting from the textual sources. We have then extended our experiments to a larger dataset to see how much influence the underlying textual source has on the result. We generally found similar trends to Wiegand et al. (2019), but much less pronounced, which indicates that the textual source has more influence on the results than the sampling strategy. Another important variable is the definition of topic: If we narrow the topic to one word, the proportion of abusive posts increases (but is still well below two of the three dataset using biased topic sampling, Waseem and Kumar), and the F-score decreases (but is still higher than all of the F-scores reported by Wiegand et al. (2019)[6]). All of our findings emphasize the importance of testing across different datasets. We have also seen the importance of having a high quality lexicon in order to determine the difference between explicit and implicit abuse.

For the future, we plan to have a closer look at the datasets since it is still unclear why some of the datasets are more difficult to classify with high accuracy than others, which cannot be explained by class skewing, sampling technique, or proportion of explicit and implicit abuse. Additionally, we will experiment with settings in which training and test data have different biases, i.e., if we sampled using different sampling strategies. We will also extend our efforts to create high quality lexicons of explicit abuse. Our ultimate goal is to improve classification performance for implicit abuse.

[6]Note than Wiegand et al. (2019) used a different classifier.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *WWW '19: The World Wide Web Conference*, pages 49–59, San Francisco, CA.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, pages 1–11, Santa Fe, NM.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2799–2804, Brussels, Belgium.

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.

Juliet van Rosendaal, Tommaso Caselli, and Malvina Nissim. 2020. Lower bias, higher density abusive language datasets: A recipe. In *Proceedings of the LREC Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 14–19, Marseille, France.

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.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.

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.

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