Reducing Gender Bias in Abusive Language Detection

Ji Ho Park, Jamin Shin, Pascale Fung
Centre for Artificial Intelligence Research (CAiRE)
Hong Kong University of Science and Technology
{jhpark, jmshinaa}@connect.ust.hk, pascale@ece.ust.hk

Abstract

Abusive language detection models tend to have a problem of being biased toward identity words of a certain group of people because of imbalanced training datasets. For example, "You are a good *woman*" was considered "sexist" when trained on an existing dataset. Such model bias is an obstacle for models to be robust enough for practical use. In this work, we measure gender biases on models trained with different abusive language datasets, while analyzing the effect of different pre-trained word embeddings and model architectures. We also experiment with three bias mitigation methods: (1) debiased word embeddings, (2) gender swap data augmentation, and (3) fine-tuning with a larger corpus. These methods can effectively reduce gender bias by 90-98% and can be extended to correct model bias in other scenarios.

1 Introduction

Automatic detection of abusive language is an important task since such language in online space can lead to personal trauma, cyber-bullying, hate crime, and discrimination. As more and more people freely express their opinions in social media, the amount of textual contents produced every day grows almost exponentially, rendering it difficult to effectively moderate user content. For this reason, using machine learning and natural language processing (NLP) systems to automatically detect abusive language is useful for many websites or social media services.

Although many works already tackled on training machine learning models to automatically detect abusive language, recent works have raised concerns about the robustness of those systems. Hosseini et al. (2017) have shown how to easily cause false predictions with adversarial examples in Google's API, and Dixon et al. (2017) show that classifiers can have unfair biases toward certain groups of people.

We focus on the fact that the representations of abusive language learned in only supervised learning setting may not be able to generalize well enough for practical use since they tend to overfit to certain words that are neutral but occur frequently in the training samples. To such classifiers, sentences like "You are a good woman" are considered "sexist" probably because of the word "woman."

This phenomenon, called *false positive bias*, has been reported by Dixon et al. (2017). They further defined this model bias as unintended, "a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others."

Such model bias is important but often unmeasurable in the usual experiment settings since the validation/test sets we use for evaluation are already biased. For this reason, we tackle the issue of measuring and mitigating unintended bias. Without achieving certain level of generalization ability, abusive language detection models may not be suitable for real-life situations.

In this work, we address model biases specific to gender identities (gender bias) existing in abusive language datasets by measuring them with a generated unbiased test set and propose three reduction methods: (1) debiased word embedding, (2) gender swap data augmentation, (3) fine-tuning with a larger corpus. Moreover, we compare the effects of different pre-trained word embeddings and model architectures on gender bias.

2 Related Work

So far, many efforts were put into defining and constructing abusive language datasets from different sources and labeling them through crowd-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2799–2804
Brussels, Belgium, October 31 - November 4, 2018.

sourcing or user moderation (Waseem and Hovy, 2016; Waseem, 2016; Founta et al., 2018; Wulczyn et al., 2017). Many deep learning approaches have been explored to train a classifier with those datasets to develop an automatic abusive language detection system (Badjatiya et al., 2017; Park and Fung, 2017; Pavlopoulos et al., 2017). However, these works do not explicitly address any model bias in their models.

Addressing biases in NLP models/systems have recently started to gain more interest in the research community, not only because fairness in AI is important but also because bias correction can improve the robustness of the models. Bolukbasi et al. (2016) is one of the first works to point out the gender stereotypes inside word2vec (Mikolov et al., 2013) and propose an algorithm to correct them. Caliskan et al. (2017) also propose a method called Word Embedding Association Test (WEAT) to measure model bias inside word embeddings and finds that many of those pretrained embeddings contain problematic bias toward gender or race. Dixon et al. (2017) is one of the first works that point out existing "unintended" bias in abusive language detection models. Kiritchenko and Mohammad (2018) compare 219 sentiment analysis systems participating in SemEval competition with their proposed dataset, which can be used for evaluating racial and gender bias of those systems. Zhao et al. (2018) shows the effectiveness of measuring and correcting gender biases in coreference resolution tasks. We later show how we extend a few of these works into ours.

3 Datasets

3.1 Sexist Tweets (`st`)

This dataset consists of tweets with sexist tweets collected from Twitter by searching for tweets that contain common terms pertaining to sexism such as "feminazi." The tweets were then annotated by experts based on criteria founded in critical race theory. The original dataset also contained a relatively small number of "racist" label tweets, but we only retain "sexist" samples to focus on gender biases. Waseem and Hovy (2016); Waseem (2016), the creators of the dataset, describe "sexist" and "racist" languages as specific subsets of abusive language.

Name	Size	Positives (%)	μ	σ	max
`st`	18K	33%	15.6	6.8	39
`abt`	60K	18.5%	17.9	4.6	65

Table 1: Dataset statistics. μ, σ, max are mean, std.dev, and maximum of sentence lengths

3.2 Abusive Tweets (`abt`)

Recently, Founta et al. (2018) has published a large scale crowdsourced abusive tweet dataset with 60K tweets. Their work incrementally and iteratively investigated methods such as boosted sampling and exploratory rounds, to effectively annotate tweets through crowdsourcing. Through such systematic processes, they identify the most relevant label set in identifying abusive behaviors in Twitter as $\{None, Spam, Abusive, Hateful\}$ resulting in 11% as 'Abusive,' 7.5% as 'Hateful', 22.5% as 'Spam', and 59% as 'None'. We transform this dataset for a binary classification problem by concatenating 'None'/'Spam' together, and 'Abusive'/'Hateful' together.

4 Measuring Gender Biases

4.1 Methodology

Gender bias cannot be measured when evaluated on the original dataset as the test sets will follow the same biased distribution, so normal evaluation set will not suffice. Therefore, we generate a separate *unbiased test set* for each gender, male and female, using the identity term template method proposed in Dixon et al. (2017).

The intuition of this template method is that given a pair of sentences with only the identity terms different (ex. "He is happy" & "She is happy"), the model should be able to generalize well and output same prediction for abusive language. This kind of evaluation has also been performed in *SemEval 2018: Task 1 Affect In Tweets* (Kiritchenko and Mohammad, 2018) to measure the gender and race bias among the competing systems for sentiment/emotion analysis.

Using the released code[1] of Dixon et al. (2017), we generated 1,152 samples (576 pairs) by filling the templates with common gender identity pairs (ex. male/female, man/woman, etc.). We created templates (Table 2) that contained both neutral and offensive nouns and adjectives inside the vocabu-

[1] https://github.com/conversationai/unintended-ml-bias-analysis

Example Templates
You are a *(adjective) (identity term).*
(verb) (identity term).
Being *(identity term)* is *(adjective)*
I am *(identity term)*
I hate *(identity term)*

Table 2: Example of templates used to generated an unbiased test set.

Type	Example Words
Offensive	disgusting, filthy, nasty, rude, horrible, terrible, awful, worst, idiotic, stupid, dumb, ugly, etc.
Non-offensive	help, love, respect, believe, congrats, hi, like, great, fun, nice, neat, happy, good, best, etc.

Table 3: Example of offensive and non-offensive verbs & adjectives used for generating the unbiased test set.

lary (See Table 3) to retain balance in neutral and abusive samples.

For the evaluation metric, we use 1) AUC scores on the original test set (Orig. AUC), 2) AUC scores on the unbiased generated test set (Gen. AUC), and 3) the false positive/negative equality differences proposed in Dixon et al. (2017) which aggregates the difference between the overall false positive/negative rate and gender-specific false positive/negative rate. False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) are defined as below, where $T = \{male, female\}$.

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

Since the classifiers output probabilities, equal error rate thresholds are used for prediction decision.

While the two AUC scores show the performances of the models in terms of accuracy, the equality difference scores show them in terms of fairness, which we believe is another dimension for evaluating the model's generalization ability.

4.2 Experimental Setup

We first measure gender biases in `st` and `abt` datasets. We explore three neural models used in previous works on abusive language classification: Convolutional Neural Network (CNN) (Park and Fung, 2017), Gated Recurrent Unit (GRU) (Cho et al., 2014), and Bidirectional GRU with self-attention (α-GRU) (Pavlopoulos et al., 2017), but with a simpler mechanism used in Felbo et al. (2017). Hyperparameters are found using the validation set by finding the best performing ones in terms of original AUC scores. These are the used hyperparameters:

1. CNN: Convolution layers with 3 filters with the size of [3,4,5], feature map size=100, Embedding Size=300, Max-pooling, Dropout=0.5
2. GRU: hidden dimension=512, Maximum Sequence Length=100, Embedding Size=300, Dropout=0.3
3. α-GRU: hidden dimension=256 (bidirectional, so 512 in total), Maximum Sequence Length=100, Attention Size=512, Embedding Size=300, Dropout=0.3

We also compare different pre-trained embeddings, `word2vec` (Mikolov et al., 2013) trained on Google News corpus, `FastText` (Bojanowski et al., 2017)) trained on Wikipedia corpus, and randomly initialized embeddings (*random*) to analyze their effects on the biases. Experiments were run 10 times and averaged.

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

4.3 Results & Discussions

Tables 4 and 5 show the bias measurement experiment results for `st` and `abt`, respectively. As expected, pre-trained embeddings improved task performance. The score on the unbiased generated test set (Gen. ROC) also improved since word embeddings can provide prior knowledge of words.

However, the equality difference scores tended to be larger when pre-trained embeddings were

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

used, especially in the `st` dataset. This confirms the result of Bolukbasi et al. (2016). In all experiments, direction of the gender bias was towards female identity words. We can infer that this is due to the more frequent appearances of female identities in "sexist" tweets and lack of negative samples, similar to the reports of Dixon et al. (2017). This is problematic since not many NLP datasets are large enough to reflect the true data distribution, more prominent in tasks like abusive language where data collection and annotation are difficult.

On the other hand, `abt` dataset showed significantly better results on the two equality difference scores, of at most 0.04. Performance in the generated test set was better because the models successfully classify abusive samples regardless of the gender identity terms used. Hence, we can assume that `abt` dataset is less gender-biased than the `st` dataset, presumably due to its larger size, balance in classes, and systematic collection method.

Interestingly, the architecture of the models also influenced the biases. Models that "attend" to certain words, such as CNN's max-pooling or α-GRU's self-attention, tended to result in higher false positive equality difference scores in `st` dataset. These models show effectiveness in catching not only the discriminative features for classification, but also the "unintended" ones causing the model biases.

5 Reducing Gender Biases

We experiment and discuss various methods to reduce gender biases identified in Section 4.3.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Table 6: Results of bias mitigation methods on `st` dataset. 'O' indicates that the corresponding method is applied. See Section 5.3 for more analysis.

5.1 Methodology

Debiased Word Embeddings (DE) (Bolukbasi et al., 2016) proposed an algorithm to correct word embeddings by removing gender stereotypical information. All the other experiments used pretrained word2vec to initialized the embedding layer but we substitute the pretrained word2vec with their published embeddings to verify their effectiveness in our task.

Gender Swap (GS) We augment the training data by identifying male entities and swapping them with equivalent female entities and vice-versa. This simple method removes correlation between gender and classification decision and has proven to be effective for correcting gender biases in coreference resolution task (Zhao et al., 2018).

Bias fine-tuning (FT) We propose a method to use transfer learning from a less biased corpus to reduce the bias. A model is initially trained with a larger, less-biased source corpus with a same or similar task, and fine-tuned with a target corpus with a larger bias. This method is inspired by the fact that model bias mainly rises from the imbalance of labels and the limited size of data samples. Training the model with a larger and less biased dataset may regularize and prevent the model from over-fitting to the small, biased dataset.

5.2 Experimental Setup

Debiased word2vec Bolukbasi et al. (2016) is compared with the original *word2vec* (Mikolov et al., 2013) for evaluation. For gender swapping data augmentation, we use pairs identified through crowd-sourcing by Zhao et al. (2018).

After identifying the degree of gender bias of each dataset, we select a source with less bias and a target with more bias. Vocabulary is extracted from training split of both sets. The model is first trained by the source dataset. We then remove final softmax layer and attach a new one initialized for training the target. The target is trained with a slower learning rate. Early stopping is decided by the valid set of the respective dataset.

Based on this criterion and results from Section 4.3, we choose the `abt` dataset as source and `st` dataset as target for bias fine-tuning experiments.

5.3 Results & Discussion

Table 6 shows the results of experiments using the three methods proposed. The first rows are the baselines without any method applied. We can see from the second rows of each section that debiased word embeddings alone do not effectively correct the bias of the whole system that well, while gender swapping significantly reduced both the equality difference scores. Meanwhile, fine-tuning bias with a larger, less biased source dataset helped to decrease the equality difference scores and greatly improve the AUC scores from the generated unbiased test set. The latter improvement shows that the model significantly reduced errors on the unbiased set in general.

To our surprise, the most effective method was applying both debiased embedding and gender swap to GRU, which reduced the equality differences by 98% & 89% while losing only 1.5% of the original performance. We assume that this may be related to the influence of "attending" model architectures on biases as discussed in Section 4.3. On the other hand, using the three methods together improved both generated unbiased set performance and equality differences, but had the largest decrease in the original performance.

All methods involved some performance loss when gender biases were reduced. Especially, fine-tuning had the largest decrease in original test set performance. This could be attributed to the difference in the source and target tasks (abusive & sexist). However, the decrease was marginal (less than 4%), while the drop in bias was significant. We assume the performance loss happens because mitigation methods modify the data or the model in a way that sometimes deters the models from discriminating important "unbiased" features.

6 Conclusion & Future Work

We discussed model biases, especially toward gender identity terms, in abusive language detection. We found out that pre-trained word embeddings, model architecture, and different datasets all can have influence. Also, we found our proposed methods can reduce gender biases up to 90-98%, improving the robustness of the models.

As shown in Section 4.3, some classification performance drop happens when mitigation methods. We believe that a meaningful extension of our work can be developing bias mitigation methods that maintain (or even increase) the classification performance and reduce the bias at the same time. Some previous works (Beutel et al.; Zhang et al., 2018) employ adversarial training methods to make the classifiers unbiased toward certain variables. However, those works do not deal with natural language where features like gender and race are latent variables inside the language. Although those approaches are not directly comparable to our methods, it would be interesting to explore adversarial training to tackle this problem in the future.

Although our work is preliminary, we hope that our work can further develop the discussion of evaluating NLP systems in different directions, not merely focusing on performance metrics like accuracy or AUC. The idea of improving models by measuring and correcting gender bias is still unfamiliar but we argue that they can be crucial in building systems that are not only ethical but also practical. Although this work focuses on gender terms, the methods we proposed can easily be extended to other identity problems like racial and to different tasks like sentiment analysis by following similar steps, and we hope to work on this in the future.

Acknowledgments

This work is partially funded by ITS/319/16FP of Innovation Technology Commission, HKUST, and 16248016 of Hong Kong Research Grants Council.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *FAT/ML 2018: 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics Volume 5, Issue 1*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP2014*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2017. Measuring and mitigating unintended bias in text classification. In *AAAI*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *EMNLP2017*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *AAAI*.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics(*SEM), New Orleans, USA*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *ALW1: 1st Workshop on Abusive Language Online to be held at the annual meeting of the Association of Computational Linguistics (ACL) 2017*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135.

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.

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.

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. International World Wide Web Conferences Steering Committee.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *Proceedings of AAAI/ACM Conference on Ethics and Society(AIES) 2018*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *NAACL 2018*.