Gender Biases in Abusive Language Detection using Transformer-based Models

Gengxiao Li
Stanford University
gli01@stanford.edu

Dingyi (Penny) Pan
Stanford University
dpan3@stanford.edu

Abstract

In this project, we evaluated the performance of different transformer-based models on abusive language detection task and measured the biases that were introduced during the detection process. In general, the transformer-based models outperform simple neural network models. With the focus on BERT, we also tested different strategies to reduce the induced biases, and the results suggest that gender swapping reduces biases more without significantly affecting the original performance on abusive language detection task. On the other hand, bias fine-tuning may be affected by the imbalanced distribution of labels that leads to an exacerbation in the model performance[1].

1 Introduction

Biases and stereotypes toward members of certain minority groups are embedded in and propagate through languages. Yet, with the inundation of posts on social media, it is impossible for platforms to manually identify and remove each harmful message. Hence, natural language understanding systems are useful in assisting the detection of toxic messages and identifying biases. Yet, the detection is not always easy. Whether a sentence should be classified as toxic or harmful depends on the speaker's identity and intention as well as the identity of the targeted group. A seemingly pejorative term may be reclaimed or used as a self-mockery. Thus, simply classifying the text by word meanings without considering these contexts will even further prevent voices of minority groups from being heard (Sap et al., 2019). Furthermore, because of the imbalanced distribution of different identity terms in the training datasets for hate speech detection, models overfit the data and incorrectly correlate these identity terms with toxicity, which introduces new biases toward the minority groups (Park et al., 2018). In addition, some biases are implicit, and the listener obtains the implicature by reasoning the intention of the speaker. Hence, this implied bias requires the model to have commonsense knowledge in order to make the correct inference (Sap et al., 2020).

In order to build models that can accurately detect toxic messages without reinforcing biases, researchers attempt to identify the source of the biases: Whether it is because of the negligence of context during annotation, or the imbalanced distribution of the data, or the structure or the learning algorithm of the model. Additionally, they propose different frameworks to understand how to detect social biases embedded in texts, how to mitigate the effects of biases, and ultimately, how we can use current technology, such as text generation, to rewrite the socially biased text.

In this project, we first examined the gender biases generated by BERT during the abusive language detection task. Then, we evaluated the effect of different strategies, specifically gender swapping and bias fine-tuning on different datasets as proposed by Park et al. (2018) in reducing the induced biases during toxic language detection.

2 Prior Literature

Park et al. (2018) focuses on biases toward certain identity groups due to imbalanced training datesets when looking at the result of abusive language detection. This type of bias is an obstacle for models to be robust enough for practical use. In this paper, the authors tested three different neural network structures, namely convolutional neural network (CNN), Gated Recurrent Unit (GRU), and Bidirectional GRU with self-attention (α-GRU), on two twitter datasets that differ in size and the degree of biases. Moreover, they also evaluate the effect of the word representation using three pre-trained embeddings, word2vec, FastText, and randomly initialized embeddings. The evaluation result showed

[1]The code to the project can be found https://github.com/pennydy/cs224u_proj.git

that using the larger less baised dataset has the significantly better result on the two equality difference scores. The performance was better in generated test set (unbiased), which suggests that the models can successfully classify abusive samples regardless of the gender identity terms.

Next, the paper introduces three gender reducing methods: 1) Debiased word embeddings, which is proposed in Bolukbasi et al. (2016) to correct word embeddings by removing gender stereotypical information, 2) Gender swap, which identifies male entities and swaps them with equivalent female entities and vice-versa, and 3) Bias fine-tuning, where the model is 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. Using dataset from different distribution can regularize and prevent the model from over-fitting to the small, biased dataset. The results show that word embeddings do not have any impact on correcting the bias. Yet, gender swapping significantly improves both the equality difference scores, and fine-tuning bias with a larger and less biased source dataset helps to decrease the equality difference scores and greatly improve the AUC scores from the generated unbiased test set. Applying both debiased embedding and gender swap to GRU gives the best performance, which significantly reduces the equality differences without losing much of the original performance. However, all methods have some performance loss when gender biases were reduced. It is possible that mitigation methods modify the data or the model, which prevents the model from accurately discriminating those critical features.

With the introduction of large pre-trained language models that outperform many simple neural network models on various benchmarks, people have shifted to the use of these transformer-based models, such as BERT. Hence, we predicted that BERT should outperform other neural networks models that were tested in Park et al. (2018) in the classification task, yet since these transformer-based models rely on attention, it is possible that they will also introduce implicit biases during detecting the abusive language use.

3 Data

In the original paper, Park et al. (2018) used two datasets, namely Abusive Tweets (`abt`) and Sexist Tweets (`st`) datasets, and to evaluate and measure biases, the authors created another unbiased dataset that contains sentence pairs that only differ in the identity terms, following the templates proposed by Dixon et al. (2018).

Since Park et al. (2018) does not provide any data source available online, we obtain the Abusive Tweets (`abt`) from the original paper Founta et al. (2018). It contains 99,799 tweets IDs, and each tweet is annotated with one of the following four labels {None, Spam, Abusive, Hateful}. Since this file does not include any text, we used twarc package[2] to connect the Tweet API and link these tweets IDs to the actual tweet contents. Once we obtained the actual tweet text for each tweet id, we merged the tweet text info to the associate tweet IDs with the corresponding label. After merging and dropping none values, there are only 30,818 tweets left with the actual text info. Then, according to Park et al. (2018), we concatenate "None/Spam" together as a "normal" label, and "Abusive/Hateful" together as an "abusive" label for a binary classification problem.

Using the similar method, we constructed the Sexist Tweets (`st`) dataset using the tweets IDs and the annotated labels as provided in Waseem and Hovy (2016). Since the current project only focuses on gender biases, we filtered out other types of biases and only keep tweets that are labeled as either not abusive or sexist.

The unbiased test set was generated using a template from Dixon et al. (2018) for male and female. The reason to develop this test set is when evaluating on the original dataset, gender bias is hard to measure because the test sets will all have the same skewed distribution, with a disproportionately large amount of abusive texts. Hence, a normal evaluation set will not be sufficient. The idea behind this template technique is that if there is no induced bias, then for a set of sentences that only differs in the identity words. For example, for the sentence pair "He is happy" and "She is happy", the model should provide the same forecast for abusive language labelling, since they are both non-abusive and only differ in the gender pronoun. By populating the templates (Figure 1) with common gender identification pairings (e.g., male/female, man/woman, etc.), we created 1,148 samples (574 pairs). To maintain balance in neutral and abusive examples, we constructed templates that comprised both neutral and offensive verbs and adjectives (Figure 2) inside the lexicon.

[2] https://github.com/DocNow/twarc

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

Figure 1: The templates used to generate the unbiased test set (Table 1 in Park et al. (2018).)

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.

Figure 2: Example of offensive & non-offensive verbs and adjectives used for generating the unbiased test set (Table 3 in Park et al. (2018).)

4 Model

For measuring biases in the abusive language detection task, we implemented pre-trained BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) model. We used the pre-trained weights of each model as provided in the HuggingFace library. In order to finetune the BERT model with our data, we added a fully connected layer on top of the output corresponding to the [CLS] token, which is commonly used as the representation of the sentence. In addition, we also implemented hyperparameter search in order to optimize hyperparameters. Since BERT had relatively better performance in the classification task and induces fewer biases than RoBERTa, we then evaluated the effects of gender swapping and bias fine-tuning using BERT.

5 Methods

5.1 Evaluation Metrics

In Park et al. (2018), AUC score, False Negative Equality Difference (FPED) and False Positive Equality Difference (FNED)[3] are selected as the evaluation metrics, and we will adopt these metrics in this project as well.

Area Under the Curve (AUC) score measures the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. High AUC score means the better performance of the model at distinguishing between the positive and negative classes. False Negative Equality Difference (FNED) or False Positive Equality Difference (FPED) are the difference between the overall false negative/positive rate and gender-specific false negative/positive rate and aggregate the absolute difference together. The higher the scores are, the more gender biases are involved in the classification task. The combination of these two metrics are great supplements. AUC score indicates the performance of the model, while FNED and FPED show the fairness of the model in another dimension.

[3] 'Not abusive' (**NOT-BAD**) is considered as the positive label in the confusion matrix.

5.2 Measuring Biases

Since the current project focuses on evaluating transformer-based models, we used the CNN, GRU, and α-GRU models from the original paper (Park et al., 2018) as our baselines. compared the results with those in the original paper. First, we tested the model on `abt` and `st` dataset respectively for the abustive language detection task and evaluated the induced gender biases. We used 80% of the data for training and assessed the model performance on the remaining 20% data using the original AUC score. Then, we also trained the model using the entire dataset and tested it on unbiased test set to obtain the generated AUC score, FNED, and FPED. The results are shown in the below section (Table 2). For the BERT model, we used the pretrained "bert-base-uncased" tokenizer as provided by Huggingface. For the RoBERTa model, we used the pretrained "roberta-base" tokenzier, which differs from the BERT tokenizer as it uses byte-level encoding. In order to optimize the set of hyperparameters for each model of each dataset, we used grid search with cross-validation. The training was stopped if there as no decrease in loss after five iterations. These are the used hyperparameters for abt and st:

1) ABT: gradient accumulation steps=8, hidden dim=200, batch size=8, lr=3e-5

2) ST: gradient accumulation steps=4, hidden dim=200, batch size=8, lr=5e-5

5.3 Reducing Biases

Next, we tested if the debiasing methods that are proposed in Park et al. (2018) can be applied to the abusive language detection model using BERT. Park et al. (2018) mentions three debiasing methods. The first method is using debiased word em-

beddings, but from the result shown in Park et al. (2018), it reduces neither the FNED nor the FPED score when used alone. We decide not to proceed with it. The second method is gender swap, which swaps the words of gender entities to the opposite gender in the training set. The goal is to supplement the training data by detecting male identities and replacing them with female counterparts, and vice versa. This basic strategy has been shown to be successful in addressing gender biases by removing the association between gender and classification choice Zhao et al. (2018). The last method is bias fine-tuning, which uses a lager, less biased dataset as the training set and a smaller, but more biased dataset with same or similar topics as the validation set to fine-tune the model parameters. The idea behind this method is that the bias in the model is mainly caused by the imbalance in the labels and the small sample size of the data. The model needs to be regularized and trained with a bigger, less biased dataset, so that it will not overfit the tiny, biased dataset.

Similar to the training-testing strategy for measuring biases, We trained the baseline with BERT classifier on combined dataset of `abt` and `st` on a 8/2 train-test split, and assessed the original AUC score. Then we trained on the combined dataset, and tested on the unbiased test set for the generated AUC score, FNED, and FPED. Next in terms of the debiasing methods, for gender swap, We trained on combined dataset of `abt` and `st` with gender swap on a 8/2 train test split, and assessed the original AUC score. Then trained on the combined gender swap dataset, and tested on the unbiased test set for generated AUC score, FNED, and FPED. For fine-tuning, we separated the `st` dataset into a 8/2 train test slipt, and 80% of the combined dataset of the `abt` and `st` train as the training set for the BERT classifier. Then we used the `st` test to fine tune the top layer of BERT classifier. The original AUC score was assessed by the 20% of the combined dataset, while the generated AUC, FNED and FPED were based on the results of the unbiased test set. For gender swap with finetuning (GS+FT) method, the process and the model structure remain the same as fine-tuning, and the gender swap dataset was used instead of the original `abt` and `st` datasets. The testing results are shown in the below section (Table 3).

6 Results

Dataset	Model	Orig. AUC	Gen. AUC	FNED	FPED
`abt`	CNN	0.956	0.999	0.002	0.021
	GRU	0.952	0.997	0.017	0.037
	α-GRU	0.955	0.999	0.012	0.026
`st`	CNN	0.906	0.635	0.305	0.263
	GRU	0.887	0.633	0.301	0.254
	α-GRU	0.890	0.631	0.315	0.306

Table 1: The AUC, FNED, FPED scores of the baseline models as reported in Park et al. (2018). All models reported here use word2vec embeddings, as the overall performance is better than the others.

Dataset	Model	Orig. AUC	Gen. AUC	FNED	FPED
`abt`	BERT	**0.972**	**0.980**	**0.000**	0.003
	RoBERTa	0.968	0.944	0.066	**0.000**
`st`	BERT	**0.927**	0.706	**0.265**	0.474
	RoBERTa	0.910	**0.713**	0.603	**0.178**

Table 2: The AUC, FNED, FPED scores of each dataset. The Orig. AUC score is obtained by testing on the original test set of the data, whereas the Gen. AUC score is obtained by testing on the generated unbiased test set. FNED and FPED are obtained by testing on the generated unbiased test set.

The results show that regardless of the model structure, the performance on `abt` dataset was better on the `st` dataset, which is in line with the finding in Park et al. (2018) that the `st` dataset is more biased than the `abt` dataset. Moreover, in comparison to the baseline results of GRU and CNN, the transformer-based models have higher AUC scores. Speficially, the results on original AUC, FNED and FPED of `abt` dataset are better than results on `abt` from Park et al. (2018), and the similar pattern is found for the `st` dataset as well.

In addition, compared to RoBERTa, BERT has better performance on both datasets in classifying the text. Additionally, as suggested by the overall lower FNED and FPED scores, BERT seems to induce less biases than RoBERTa.

Method	Orig. AUC	Gen. AUC	FN-ED	FP-ED
Baseline	0.952	**0.918**	0.042	0.164
Gender swap	**0.963**	0.905	**0.000**	**0.014**
Finetuning	0.767	0.763	0.010	0.251
GS+FT	0.787	0.750	**0.000**	0.017

Table 3: The AUC, FNED, FPED scores of BERT with different bias-reducing methods.

7 Analysis

To start with, both transformer-based models and simple neural networks performed really well on the classification task, since the original AUC scores are around 0.95 when being tested on the less gender-biased `abt` dataset and slightly worse when tested on the generated unbiased dataset. For the smaller yet more biased `st` dataset, Transformer-based model outperformaned the baseline neural network models on both the original testing data and the generated data. These pattern confirm our initial hypothesis about the advantage of using transformer-based models in the classification task.

Furthermore, the transformer-based models also seem to generate less biases when classifying texts in the `abt` dataset, whereas BERT produces a larger false positive rate difference (0.265) between the classification result of the two genders, indicating by the FPED score than RoBERTa (0.603), whereas the reverse is true for the false negative rate difference (0.474 for BERT and 0.178 for RoBERTa), although the best score from each of the model in each category outperforms the results of the neural networks (around 0.30 for both FNED and FPED regardless of model type). This seems to suggests that BERT-like models introduce more biases when the dataset is small and biased.

Table 3 shows the results of the debiasing methods. Overall, the most appealing one is to only apply gender swap method to the model, as it reduced the false negative equality difference and false positive equality difference by 100% and 91.5% while losing 1.4% of the original performance. We also noticed that when performing the finetuning method that is proposed by Park et al. (2018), the false positive rate, male false positive rate and female positive rate are all close to 1.0. This is disguised by the fact that FPED is close to zero, which seems to suggests that the method has successfully mitigate the effect of biases, because when all the false positive rates are close to 1.0, their difference is close to zero. According to the results shown on the confusion matrix, the positive is defined as non-abusive (**NOT-BAD**), while negative is defined as abusive (**BAD**). Hence, a close to 1.0 false positive rate means most of the abusive language are classified as non-abusive language when applying fine-tuning method. Although it did reduce false negative/positive equality difference, it greatly hurt the purpose of the model use to detect the abusive language. This fine tuning method is reducing biases for the sake of reducing biases, not incorporate enough on the original model purpose. The reason why this happened is because the non-abusive label of fine tuning dataset contains almost 3 times more data of the abusive label (Table 4). When we finetune the model on this dataset, the model parameters tend to inherit more information on non-abusive features. This leads to the result which the model is easier to predict text as non-abusive, which is positive in our scenario.

Label	Text
BAD	546
NOT-BAD	1454

Table 4: The distribution of the label (Abusive is labeled as **BAD** and Non-abusive as **NOT-BAD**) in the finetuning dataset, which are 20% data from the `st` dataset generated for testing.

Last but not the least, although the results suggest that gender swapping can reduce the biases that are generated during the detection tasks, it is crucial to point out that these are merely mitigation methods that do not prevent the model from producing these implied biases in the first place. Previous studies suggest that biases are rooted in the data, even at the early stage about how the annotators interpret and label the data, and biases are context-dependent. Specifically, for bias identification, Sap et al. (2017) and Sap et al. (2020) provide frameworks that attempt to classify and represent biases. The former shows that different priming methods changes the way annotators categorize the seemingly toxic texts, especially those that are related to racial information inferred from various the dialectical use. The latter provides a more general framework about how to help the model make pragmatic inferences about biases and harmful messages. Similarly, Zhou et al. (2021) and Xu et al. (2021) point out that it is the inherent biases in

the data that causes the poor performance of toxic language detection and detoxification methods on texts that contain minority identities issues, and the both propose solutions to reduce the biases by filtering or transforming the data.

Hence, more effective bias-reducing methods should adopt different annotation strategies and use common-sense knowledge to reason the source of the abusiveness in the language and the potential target group.

8 Conclusion

In general, for this project we used different transformer based models to detect abusive languages. Overall, the performance of these models are better than the neutral network based models. Also we tried different debiasing methods with BERT model, it turns out the gender swap combine with BERT model can give us the best performance on AUC score and equality difference, and it also outperforms the best combination on neutral network model with debiasing methods in Park et al. (2018).

Yet, after applying debiasing methods, most of models still have really high false positive rate, which suggests that the model tends to label most texts as non-abusive, regardless of the gender. This may lead a smaller FNED score, and the model seems to be less biased, but the overall performance in the classification task. Future studies can focus on how to balance this trade-off between mitigating biases and maintaining the task performance.

In addition, in the project, we focused only on evaluating two debiasing methods using BERT and the `abt` and `st` datasets, and in the future we plan to other more balanced datasets to test the effect of these debiasing methods with other transformer-based models, such DistilBERT (Sanh et al., 2019), which is a lot smaller than the BERT and RoBERTa model and can reduce the training time.

Known Project Limitations

1. One limitation of this project is the dataset dependency. All the debiaing methods we performed are highly dependent on the dataset we used. Especially for finetuning method, since there are more data labeled as non-abusive, more predictions of fine-tuned model were labeled as non-abusive. On the other hand, if the finetuning dataset contains more data with abusive language use, the predictions will tend to label more text as abusive, which may resulted in higher false positive rate.

2. Another worth-noting point is the goal dependency. The original purpose of this model is used to detect abusive languages, but after applying the debiasing methods, we found out that the performance of all models has decreased in different degrees. One thing always need to keep in mind is that the goal of the project. If the goal is to try to detect all the abusive languages without missing any, while still debiasing the model, then using these datasets in combination with the debiasing methods, such as gender swapping, finetuning, may not be a optimal solution for this situation. This combination will push the false positive rate close to 1.0 (see details in above section). It might need to change the dataset which contains more abusive labels in order to obtain the project goal with debiasing.

Authorship Statement

All authors contributed equally to this project. Gengxiao Li collected and preprocessed the data. Dingyi Pan helped set up the baseline and transformer-based models. Both of us worked on training the models and analyzing the results. All papers were written jointly.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv e-prints*, page arXiv:1607.06520.

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

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

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. In *Twelfth International AAAI Conference on Web and Social Media*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

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 *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

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, California. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. 2021. Challenges in automated debiasing for toxic language detection. In *EACL*.

A Example Appendix

A.1 Baseline models

We attempted to reimplement the neural network models that were similar to those tested in the original paper (Park et al., 2018), namely CNN and GRU with Glove embeddings. Specifically, we randomized the data and split them into the train/dev/test set with 8/1/1 ratio. Then, we evaluated the performance of each model on the two datasets and on the unbiased test set as well.

Dataset	Model	Orig. AUC	Gen. AUC
abs	CNN	0.948	0.045
	GRU	0.937	0.051
st	CNN	0.769	0.374
	GRU	0.531	0.303

Table 5: The AUC scores of each model. The Orig. AUC score is obtained by testing on the original test set of the data, whereas the Gen. AUC score is obtained by testing on the generated unbiased test set.

However, the results do not match with those from the original paper (Table 1). These discrepancies can be attributed to several factors. First, as mentioned in the Data section, we did not have access to the exact same datasets as the original paper, and the datasets that we constructed using the tweets IDs and twitter API contain fewer data (30k and 10k in the abs and st dataset, respectively, in comparison to 60k and 18k data used in the original study). In addition, there may be some problems with the implementation of the models.