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