Paper Title:
Detecting Unintended Social Bias in Toxic Language Datasets

Paper Link:
https://aclanthology.org/2022.conll-1.10/

1 Summary

1.1 Motivation

This paper states the importance of research on detecting unintended social bias in toxic language datasets. Here, the use of social media is increasing day by day and the rise of online hate speech and online toxicity and its automation to detect hate speech and online toxicity are becoming very famous. But in reality, very little research has been done in this field, especially in this dataset. The writers of this paper intend to fill the gap in the research. There is less than a little research on this topic of automatic detection of hate speech and offensive texts and comments.

1.2 Contribution

The first contribution of this paper is the introduction of a new dataset named ToxicBias which has a few instances that were annotated as five different bias categories namely, gender, race/ethnicity, religion, political, and LGBTQ. This is also the first study of this kind where they extract social bias from the toxic language datasets. There were also studies of model bias discussion and its countermeasures.

1.3 Methodology

The methodology of this paper is dataset creation. The writers created the ToxicBias dataset by curating different instances from existing datasets from a competition in Kaggle named "Jigsaw Unintended Bias in Toxicity Classification". They chose the instances that have toxic language and annotated them for five different bias categories: gender, race/ethnicity, religion, political, and LGBTQ. The authors also, trained transformer-based models using the curated ToxicBias dataset.

1.4 Conclusion

In this paper, the author demonstrated that the identification of attacks and hate speech online is often assimilated with social biases or stereotypes. Nevertheless, all hate speech and attacks are not based on social biases and some are intended only on personal hate. In addition, detecting bias without any prior context is also proven to be tough for the annotators and as well as the models to detect. To conclude, this paper proves that, biases can also have different directions. Biases can occur both in favor and against a community.

2 Limitations

2.1 First Limitation

The first limitation of this paper is the lack of external context and small-sized datasets. In this dataset, there is no inclusion of external datasets which is very inaccurate for the categorization task.

2.2 Second Limitation

There are only five types of social biases that are considered in this research. There is much more than these five categories than these social biases. Also, this research is based on the English language and the dataset is oriented towards Western culture. Any non-western culture wouldn't match with these datasets. Also, there are no dealings with multilingual biases.

3 Synthesis

In this paper, the author proposes and issues the critical necessity of detecting unintended social bias in toxic language datasets. This field has received very little attention in the field of automation and natural language processing. Here, the writers introduced a new dataset called ToxicBias which was derived from a competition on Kaggle. This paper involves dataset creation, annotation, and model training with detailed guidelines to unleash the harmful biases of toxic language datasets. This paper brings a shedding light on the important aspects of NLP and provides curated datasets with identification and mitigation. The dataset ToxicBias can benefit future researchers.