CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Arjun Neervannan**	**Project Number** **S0415**

Project Title

Combating Cyberbullying and Toxicity by Teaching AI to Use Linguistic Insights from Human Interactions in Social Media

Abstract

Objectives

Social Media is replete with toxic comments. AI algorithms built to identify toxicity often exhibit bias as they associate identity terms with toxicity, lacking an understanding of context. Current approaches to bias-free toxic classification in forums do not scale due to manual identity term selection and un-interpretability. Automatically detecting the identity terms associated with toxicity, and surgically removing those biases by adding more non-toxic comments containing those terms will help to make a more accurate and less biased toxic classifier. Thus, the objective was to develop a targeted AI algorithm to scale the identity term identification in removal of bias and to improve toxic comment classifiers.

Methods

Linux computer with GPUs, Python 2.7, Python libraries, such as Keras, Tensorflow, Pandas, Spacy. Since bias primarily arises from over-representation of identity terms in toxic comments, prior papers debiased toxic classification models by manually adding non-toxic comments with manually selected identity terms. Also, many identity terms in social media were used as nouns and adjectives, leading to the noun-adjective rule. This project used an Hierarchical Attention sequence learning neural network to show what terms influenced toxic classification and applied the Noun-Adjective filter to automatically detect large number of identity terms. Though adding non-toxic comments helped debias, there existed an optimal number of non-toxic comments that had the largest impact on the performance. Thus a grid search across the number of comments and percent of identity terms was run to determine the optimal point. The Area Under Curve (AUC) was used as a measure of accuracy and the False Positive Equality Difference (FPED) Improvement, which measured equity of model performance across identity terms, was used as a bias metric.

Results

The model achieved an AUC of 0.98, compared to 0.95 in a prior Google paper and achieved an FPED Improvement of 44%, with 200+ identity terms to debias compared to 50 terms in prior paper.

Conclusions

The results supported the hypothesis; the FPED Improvement of 44% supported that the automated model was able to scale better to fix the bias on 200+ terms with a high accuracy. Unlike prior papers, this model did not have comment length limitations, was language agnostic and automated, thus likely extensible to other languages.

Summary Statement

Devised a less biased, scalable AI algorithm to combat cyberbullying in social media by using linguistic insights from human interactions in social media to identify toxic comments better than prior approaches without curbing free speech.

Help Received

I designed and implemented the model myself after understanding the Hierarchical Attention Network. I reviewed results with Prof. Sameer Singh, Computer Science Department at UCI, received guidance on improving key aspects of the project, and implemented best practices when validating my methodology.