Using Natural Language Processing and Linguistic Insights to Combat Cyberbullying and Toxicity in Social Media

Neervannan, Arjun (School: University High School)

With the proliferation of social media, hateful or toxic content and cyberbullying in online forums have only grown. AI-driven algorithms deployed to moderate and reduce hateful content often exhibit bias, associating racial, gender, and other identity terms with toxicity and unduly censoring productive discussions. The censorship undermines trust and deters users that use these terms in genuine non-toxic ways. Bias-free or less biased AI models that moderate better by classifying toxic comments fairly can regain the trust of the users and facilitate safer online discussions. Current debiasing approaches are limited in their scope, not scalable due to manual feature selection, and not interpretable. While the method used in last year's paper accomplished those objectives, it introduced many noisy terms, which caused the model to overcorrect its biases and reverse some of the debiasing process. This paper is a follow-up on that. The novel method used in this project addressed these issues by using an interpretable hierarchical attention-based neural network model, using the FastText embedding network to reduce noise, adopting a linguistic-driven criteria that increased scope of important identity terms, grid-searching the metaparameters, and finally de-biasing by augmenting the dataset with the appropriate number of counter-examples. The method achieved accuracy (AUC) of 0.98 and identified several hundred more identity terms than a prior paper, while also reducing the number of noisy identity terms when compared to the GloVe model and debiasing significantly better when compared to the control set. Unlike prior approaches, this model does not have comment length limitations, while being more scalable with its use of attention networks to select and debias identity terms.