Team: LiuProvincial

Project Title: Cyberbullying Classification

Project Summary:

The importance of fostering diverse and inclusive online user communities has become increasingly pronounced in recent times due to civil unrest, COVID-19, military conflicts and economic depression. Toxic online spaces where users feel uncomfortable in expressing themselves lead to further marginalization of already marginalized groups and deter meaningful engagement amongst users, driving them away from the websites themselves. Perhaps one of the most effective means of curbing such undesirable behavior is with the preemptive identification and removal of abusive comments, which prevents any damage from being done in the first place. However, overzealous models which spuriously remove innocuous comments may also produce the same effect, as users will nonetheless feel marginalized from the perceived censorship from the website. These difficulties may suggest why large social networks such as Twitter are reluctant to implement such technologies for fear of alienating their core user base who view twitter as a means of self expression. Our goal is therefore the research and development of a precise and performant toxic Tweet classifier that exhibits high recall and precision. Toxicity can be separated into more specific classes [1]. A novelty in our challenge is to classify specific types of cyberbullying (age, ethnicity, gender, religion, and other), similar to toxicity.

Approach:

Based on preliminary research, there is only a very recent (2021) approach to binary classify cyberbullying [2]. Most pre-existing notebooks using this dataset provide simple baselines and demonstrations of how to access the data. For our baseline, we will use a simple bag of words approach followed by fully connected layers and activations. We will also experiment with training our own embeddings with convolutional neural networks as well as varying the depth and breadth. Finally, we will use more modern and trendy approaches with large pretrained language models such as the BERT and GPT families of models

Resources/Related Work:

[1] https://arxiv.org/pdf/2106.04511.pdf

[2] Gencoglu, O. (2021). Cyberbullying detection with fairness constraints. IEEE Internet Comput., 25(1):20–29

Datasets:

Cyberbullying Classification:
https://www.kaggle.com/andrewmvd/cyberbullying-classification

Team Members:

Bojun Yang

Lucas Liu