Project Proposal

Project Title: Toxic Comment Challenge

Project Teammates: Monica Bernard and Andrew Stewart

Main idea of the project: This project will focus on studying negative online behavior such has comments that are profane, vulgar, offensive, rude, disrespectful etc. These types of comments often create an unhealthy environment that can make a person leave a discussion or in extreme cases even cause hate.

In this project, we will be working on a Kaggle competition where we are challenged to build a multi-headed model that's capable of detecting different types of offensive language. These include words of threat, obscenity, insults and identity-based hate. This application can be used in several online and social media platforms such as Facebook, Twitter, Instagram, Reddit, Quora etc. There are some publicly available models served through the perspective API, including toxicity. However, the current models still make errors.

Data and Technique used: In this project, we will be using a dataset of comments from Wikipedia's talk page edits. This dataset contains large number of Wikipedia comments which have been labeled by human raters for toxic behavior and the text may be considered profane, vulgar and offensive. The types of toxicity (y labels) are toxic, severe_toxic, obscene, threat, insult and identity_hate. We will be creating a model which predicts a probability of each type of toxicity for each comment. There are several Machine Learning classifiers such as Naïve Bayes, SVM, LSTM, CNN etc., that can be used for this project. Our idea is to research and find the best model that would provide a good accuracy and implement it in Python.

Project Goal: We want to understand the dataset at hand by conducting some initial data analysis and then create a model that predicts the probability of toxicity for each comment. We expect to have a model that has above 95% accuracy, hopefully help online discussion become more productive and respectful.