Second Capstone Proposal: Toxic Comment Classifier

Dovid Burns

Toxic comments online are both damaging to people emotionally and prevent productive discussions about sensitive subjects. Many online platforms that allow for user comments such as Facebook, YouTube, Twitter, Wikipedia, Yelp and Instagram have difficulties ensuring that conversations are taking place in an appropriate way. This project will build a classifier using a dataset that contains comments from Wikipedia's talk page edits to classify them as toxic or benign. The model can then be used in many platforms to automatically detect—and possibly remove—these comments before they offend users or deter users from engaging in communications. Additionally, this can be used as a tool to find problematic users for serious warnings or banning from the site.

This data set contains almost 160,000 comments of which more than ten percent are tagged toxic. The data set to be used is publicly hosted at https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge . To create the classifier, the comments will be cleaned by first removing all non-words and then removing common English stop words. The words will then be stemmed and vectorized for analysis. Additionally, analysis on the comment length and percent of comment's letters that are capitalized will be determined from the data and binned for additional predictive features. We plan to use both a random Forest Classifier and Naïve Bayes Classifier to predict if a given comment is toxic. The deliverables will be Jupyter notebooks with python code, a full report of the analysis and a blog post on LinkedIn.