Group 9 Project Proposal Document

Scott Huff, Caleb Dease, Sean Robinson
Friday, March 16, 2018
DSBA 6156: Applied Machine Learning

Subject Domain (What is the problem, Who cares, Why should it be solved with data)
Inspired by headline news scandals and ‘fake news’ claims escalating from the 2016 era, our group finds that the domain most starving for a data-centered approach is politics. In the political environment, public opinion is one of the most critical components. But, what if a outside influence could sway those opinions with a flock of bots or henchmen behind a machine spouting rhetoric. How can humans be expected to determine the validity of an intangible persona when all of our senses are adapted to comprehend a physical environment? To combat fake news, or in our case ‘trolling’ tweets, we hope to build a model to demystify the anonymity of twitter users and their online messages. By creating a predictive model for ‘the truth’ behind the existence of a belief we can positively influence the social impact of online forums. In a more measurable outcome, a model with this scope can help news sites and political parties to more efficiently use resources. The hope is that our model will act like a computer’s antivirus program. It will be able to work in batches or anytime you want to view a tweet. It will be able to adapt based on new ‘troll’ tweets are they are uncovered. It is likely impossible to block and delete these malicious actors out of existence, so we must adapt and provide better solutions.

Data Gathering
The data for our project will come in two forms which we will later combine into our final dataset. The first of these two is a collection of Russian troll tweets (https://www.kaggle.com/vikasg/russian-troll-tweets) containing 200,000 confirmed russian tweets which contain features such as tweet text, retweet count, and favorite count along with other corresponding user data such as follower count and status count. This data is in CSV format and was gathered by NBC. While these tweets will serve as the target in our analysis, we need something to compare them to. Therefore, for our second dataset we will use the Twitter API to pull down a similar number of real tweets/user data which store them in a CSV format. First we will combine our corresponding tweet and user data by joining them on user_id to create respective real and troll datasets. After which we will clean the tweet text, add our NLP generated features, and add an is_troll class variable we will randomly combine together these four datasets to create one dataset which we can use to predict whether or not a tweet is fake.

Analysis
The analysis phase of predicting twitter bots will consist of two parts: unsupervised learning on the tweets from known bots to derive features, and using the derived features along with pre-existing features to classify the bots later on. The unsupervised learning will consist of Natural Language Processing to identify common sentiments and keywords used by the bots, and from that we will hopefully be able to gather additional meaningful features for a classification model. After the unsupervised learning has been performed on the bot tweets, a similar analysis will be ran on the tweets pulled via api and feature values will be calculated for the real tweets. Because we know the tweets which came from bots, the supervised learning phase of the bot

tweets and real tweets, each will be added to a dataset and given a class. The instances will be assigned their derived features from the unsupervised learning and user data. Multiple models will be created such as naive bayes, logistic regression, KNN, and a decision tree to find the most accurate classifier. The classifiers will be scored based on recall primarily because it is more important to minimize the false negatives for this use case. However, ROC curves can also be generated after testing the different models to look at precision of the true positive and false positive rates along with calculating the area under the curve.