Elinor Poole-Dayan
January 28, 2020
MAIS 202
Deliverable 1

I chose this dataset from Kaggle which contains almost 3 million twitter posts from the IRA controlled accounts (which are known to be Russian trolls) between 2012 and 2017. I thought this dataset was good because it contains a lot of information about the tweets in addition to the actual text of the post, including all of the tags, whether or not it was a retweet, the date and time, how many times the account was retweeted and liked, etc. I thought it would be really interesting to see if I could build a model to learn the patterns and commonalities between the most popular tweets and then maybe see if it could either generate its own fake tweets or flag other posts as troll-like. For the latter, I would need to possibly find more data on non-troll tweets or scrape data from certain twitter accounts to run the model on.

In terms of methodology, first I would need to preprocess the data. A significant number of the tweets are in Russian. Those tweets would either need to be handled separately or taken out for this project because I am not familiar with the Russian language and they could not be analyzed while mixed in with the English tweets. While excluding them altogether could introduce bias, it is fair to only look at the English tweets because the main objective of the troll accounts was to influence American politics and the vast majority of Americans also don't know Russian. Furthermore, certain columns are of especially of interest, such as the twitter author, content of the tweet, number of followers, activity on the account following the post, location, language, type (retweet or not), and account category. Since not all of these are the same data type, it might be necessary to convert categorical values to binary using one hot encoding or a similar process to make it easier for the model to interpret correctly.

This project would be carried out using the classification model. For the actual processing of the tweet content itself, I could use algorithms like tokenization or stemming to break down sentences and take out unnecessary words and inflections that don't influence the meaning of the tweet. This could also help identify common words in the troll tweets without having to deal with words that are commonly used in English in general. If there are spelling mistakes, these will have to be cleaned up. It might also be interesting to weight the hashtags or classify them separately. I could also use the Bag of Words method with TF-IDF to identify important words belonging to troll tweets and see which ones affect the classification the most.

For the application, I would like to make a web app. I am not really familiar with web app technologies, so I am not sure which one I'd use. I am thinking that it could allow the user to input a sample tweet in a text box, possibly with the option to enter a handle (if that ends up being important in the model) and then it would return whether the tweet is similar to a troll tweet (maybe output a percentage of similarity?).