Twitter data analysis for 2020 Election Prediction

Group Members: Minh-Tuan Nguyen , Aliya Alimujiang

Introduction:

Several works have shown the potential of online social media, in particular platforms like Twitter, for analyzing the public sentiment in general has a high impact on predictions. With the increasing importance of Twitter in political discussions, a considerable number of studies also investigated the possibility to analyze political processes and predict political elections from data collected on Twitter.

For our analysis, we will be using twitter data directly scrapped from twitter using R. The daily tweets are from the following hashtags: biden2020, BidenHarris2020, trump2020,MAGA, vote. The goal is to classify the content of the tweets into Trump, Biden based on the tags as well as sentiments: immigration, economy, stimulus, tax, covid ,fake news ,racist, environment, Russia, security, etc.

By classifying at the tweet level, we can correctly take into account the difference of activity of supporters to extract the percentage of users in favor of each candidate/party. This approach allows us to correctly interpret the Twitter opinion trend as the result of the variations in engagement of the supporters of each campaign and to gain unique insight on the dynamics and structure of the social network of Twitter users in relation to their political opinion.

The goal of this analysis is to be able to predict the election results solely based on twitter data to figure out the impact of twitter as well as the prediction accuracy. We will be comparing our results with the real election results at the end to conclude and verify the role that twitter plays in terms of election.

Election Proposal (Training Set & Basic Descriptive)

Aliya Alimujiang and Minh Nguyen

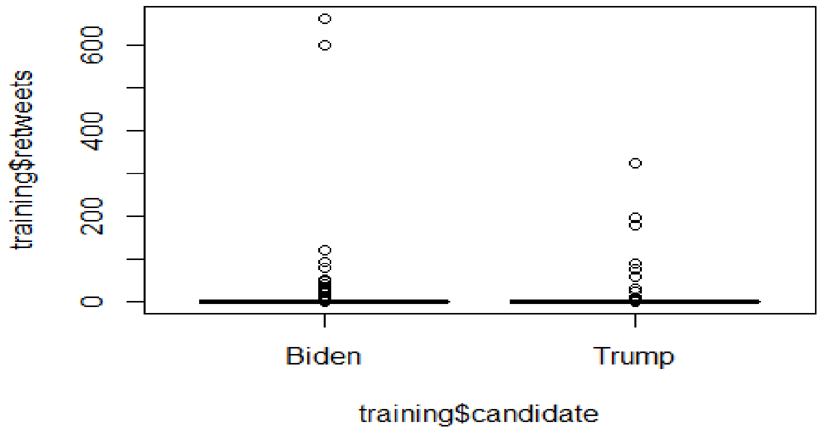
10/18/2020

Data: Live data from twitter extracted on 10/18/2020. We will be extracting data for a week for our training set. We will use another week of extracts as our test set.

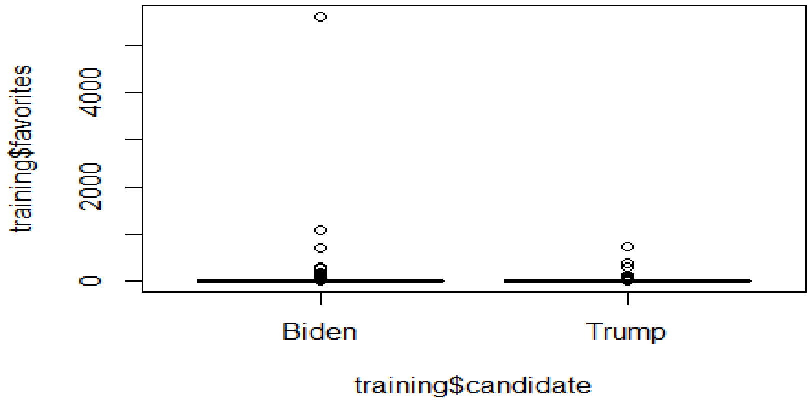
We will be first doing descriptive analytics to analyze as well as visualize the dataset. Look for correlations and if there are any violations. We will also be using variable selection process to decide which variables to keep in our final model. We will also be doing model selection process.

So far, we have collected 2 days of data from twitter. We have done a preliminary data cleanup work in R and ran some basic analysis. Based on this it looks like Biden is getting more retweets and favorites.

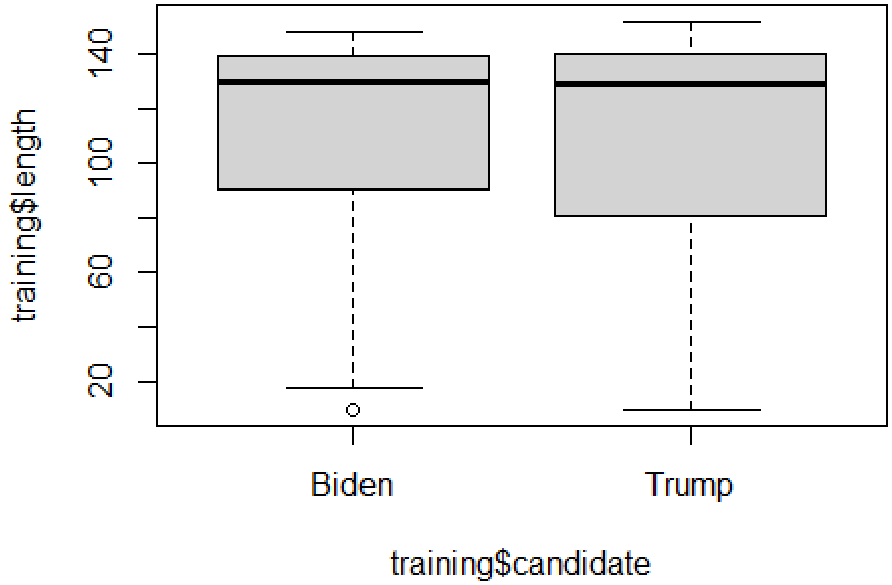
```r
training <- read.csv("../../Election and Tweets/10172020training.csv")

boxplot(training$retweets ~ training$candidate)
```



```r
boxplot(training$favorites ~ training$candidate)
```



```r
boxplot(training$length ~ training$candidate)
```



```r
training %>%
  filter(candidate == "Trump") -> df.trump

mean(df.trump$retweets)

## [1] 1.573171

mean(df.trump$favorites)

## [1] 4.171951

training %>%
  filter(candidate == "Biden") -> df.biden

mean(df.biden$retweets)

## [1] 24.95658

mean(df.biden$favorites)

## [1] 13.21456
```