DS8006: Lab 4 "Sentiment Analysis with Twitter Data"
Student's name: NAJLIS, BERNARDO (#500744793)

Note: *For this analysis, I decided to use the Election dataset (provided in the Excel spreadsheet) as both training and test data set using random sampling in splits of 70% / 30% for each subset respectively. Also, "neutral" tweets were removed from the training and test dataset. Predictions done on the twitter "Toronto" dataset were made using this same model.*

1. Briefly explain how you modified the scripts in Step 1 to load the provided training dataset.

The changes were major as the datasets provided in the example are very different from the training data set used. First this requires the use of an R library that can read Excel files ("xslx"). Then I used the "sum" field (difference between positive and negative terms in each tweet) to label positive (sum >0), neutral (sum = 0) and negative (sum < 0) tweets, and remove all neutral tweets. Then using sample() I split the set into training and test using random sampling to take 70% of all documents as the training set.

2. How accurate was the resulting prediction (recall vs. precision) ?

Here is the Confusion Matrix obtained on the test data (30% of Election dataset) and the recall / precision calculations:

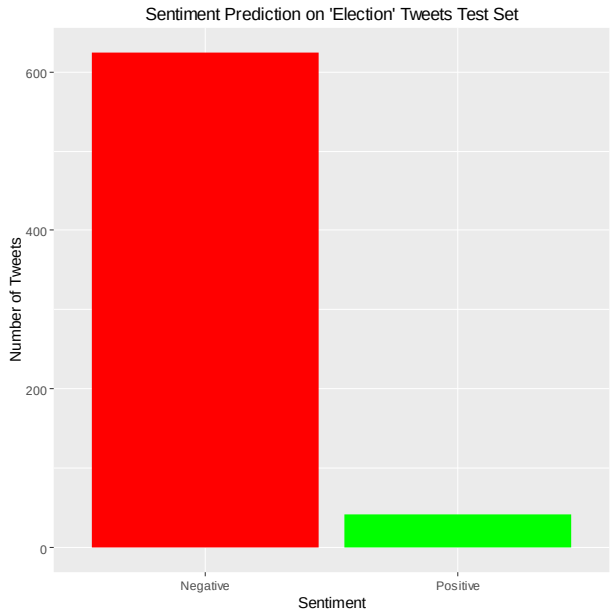
	Negative	***Positive***
Negative	***True Positive = 223***	***False Positive = 20***
Positive	***False Negative = 402***	***True Negative = 21***

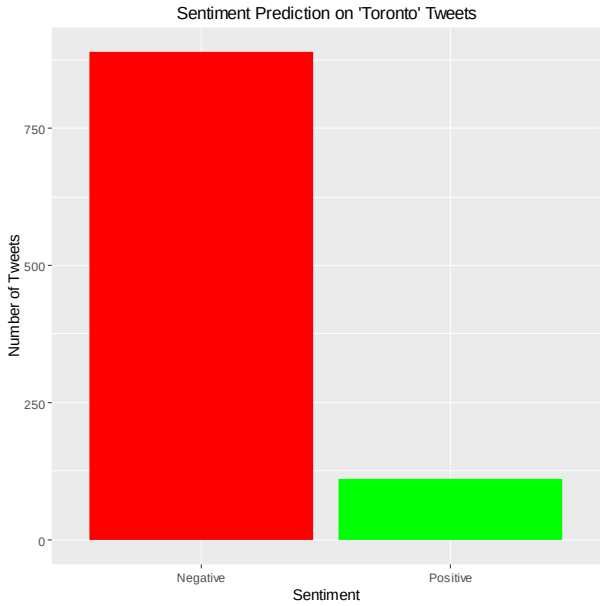
$$\textbf{\textit{Precision}} = \frac{\textbf{\textit{True positive}}}{\textbf{\textit{True Positive + False Positive}}} = \frac{\textbf{\textit{223}}}{\textbf{\textit{(223 + 20)}}} = \textbf{\textit{0.9176}}$$

$$\textbf{\textit{Recall}} = \frac{\textbf{\textit{True positive}}}{\textbf{\textit{True Positive + False Negative}}} = \frac{\textbf{\textit{223}}}{\textbf{\textit{(223 + 402)}}} = \textbf{\textit{0.3562}}$$

3. Include a screenshot of the resulting visualization.

As the prediction were run in both the test data set and the Toronto twitter dataset, here are both visualizations of number of positive and negative tweets predicted by the model.





4. Briefly explain why someone would be interested in your results.

The results can use a series of pre-classified tweets to predict sentiment on a new, unseen, data set. The precision shows the model has a higher predictive power for negative tweets (precision) than for positive tweets (recall). Also, this can be confirmed by both charts that show the number of predictive tweets is much higher than the predicted positive tweets.

5. What was the most challenging part of this lab?

The most challenging part was data preparation, taking the Excel data set and splitting into training and test.