A Federal Election Sentiment Analysis

A Makroo

Abstract

The data set included a set of sentiment classified tweets text file and unclassified political tweets during the 2015 Canadian Election. The question to be examined pertained to observing the link between public twitter sentiment with regards to political parties during the election and their eventual performance. The method included training a Naïve Bayes Classifier through Sci-kit learn library for sentiment classification. The findings show that there is a link between twitter discussions sentiments on a party and its performance, with the winner of the election having highest percentage of positive tweets from the sample.

Motivation

I’d like to understand how to predict general public sentiments about important issues, in this case particularly political parties during elections. Could an analysis of twitter data which tends to be unfiltered thoughts of the public, provide insight into the likelihood of a political party performing well during an election?

Such an analysis could give hints into the projected political success of a party, and also provide interesting comparisons with other polling techniques typically used during elections. This insight would be valuable to political analysts and general population in assessing how the rest of their citizens feel about specific political parties.

Dataset(s)

The data set I will be using is a collection of tweets during the 2015 Canadian Federal Elections, obtained through an academic institute online (University of Toronto). The data is roughly 3000 unclassified tweets pertaining to Canada during the Federal Election day (before any results were announced), and pertains to the elections that were occurring. The data is raw text (.txt) with each line in the text pertaining to a pertinent tweet.
Further, a classified generic tweet text file was also obtained, with general tweets classified as negative sentiment (0) or positive (>0), with 200,000 tweet records for training.

Data Preparation and Cleaning

The Data Preparations steps involved:

- Ingesting the txt files into pandas dfs (data frames)
- For classified tweets, extracting the first character corresponding to 0:Negative and 4:Positive
- Both set of tweet txt files dfs cleaned for punctuation and stopwords (obtained from nltk and string packages)
- Stored as two fields, tweet text and sentiment in the pandas dataframes
- Major problems were with the corner cases of cleaning the data, as this is important for feature generation for training

Research Question(s)

The intent of this project is to train a classifier on tweet texts for positive and negative sentiments, and apply it on the unclassified federal election tweet corpus to learn about public opinion of major political parties participating in the election.

The question to be addressed is what was the public sentiment in terms of positive/negative tweets for the major parties on the day of the election? How does this relate to the overall results of the elections announced later?

Methods

Ingestion: In Pandas data frames (for ease of managing text data and cleaning)

Analysis: Sci-kit learn Vectorizer was utilized to generate feature vectors for Classifier training, this allows for token generation and feature vectors generation as well. This was combined with a Naïve Bayes Classifier to train for positive and negative sentiments. This classifier was then used to predict sentiment of political tweets.

A word based algorithm was used to classify the tweets into the major parties: Liberal, Conservative, NDP and Others

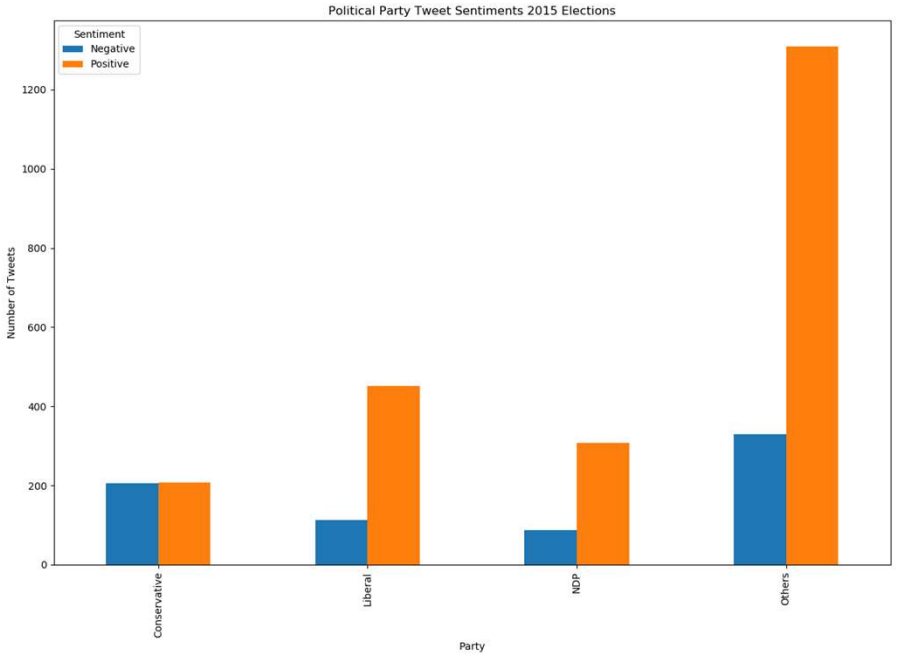
The resultant political tweets were classified for party and sentiments and graphed on matplotlib for visualization and interpretation

Findings

The results of the analysis are summarized in the following graphs:

We can see the total number of positive and negative tweets corresponding to each party in the data set (raw numbers)

This shows which party was being most talked about. Note: Others is aggregate that is other smaller parties or unclassified parties

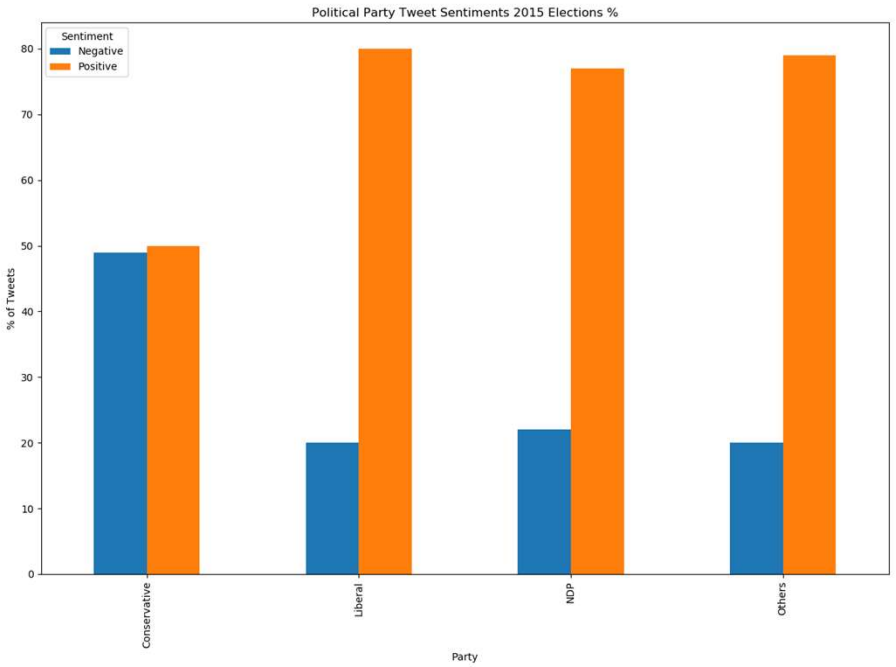


Findings

We can see the % number of positive and negative tweets corresponding to each party in the data set (normalized)

This helps provide a more comparative and scaled assessment between the parties

These results show the twitter public sentiment at election time



Limitations

Major limitations pertain to the sample size of the data set, as a comprehensive data set of tweets on election day is difficult to obtain. Also number of likes and retweets of each tweet and status of poster could be evaluated for detailed trends.

Further classification of data into political parties can be improved based on keyword terms pertaining to each party.

Conclusions

The outcome of the election was a Liberal Party win in 2015

It can be seen that the Liberal party tweets counts was found to be the most numerous, as well as the highest % of positive sentiment tweets, the conservative party which favoured poorly in the election, had a much higher ratio of negative tweets.

This demonstrates that a sentiment analysis of political tweets can be correlated with the party's performance and provide indicators to its eventual performance. The more comprehensive the collection of tweets, the better chance of successful indicators.

Acknowledgements

I collected the data myself from an academic institution

I had no feedback other than using lessons from the Mini project to improve upon in this analysis

References

All the work conducted was original and conducted by myself through aid of course materials.