Twitter Sentiment Analysis App Tracks The Pulse of India's Elections

Introduction

Twitter has declared the 2014 Indian General Elections is a "Twitter Election". Indian Elections 2014 have completely redefined the way India has witnessed political battles fought between individuals and parties. Election saw 56 million election-related Tweets during 5 months, when the polls ended. Eventually it provides whose fortunes are rising and falling in the eyes of the web, and to learn who the web thinks will be the overall nominee.

Here, candidate of main 3 parties are considered for the analysis. Candidates are

- Narendra Modi- Bharatiya Janata Party(BJP)
- Rahul Gandhi- Indian National Congress(INC)
- Arvind Kejriwal- Aam Aadmi Party(AAP)

Twitter became the medium of choice for people to engage in and consume political content. Take any metric: original content generated, engagement by political leaders, user engagement with content, news breaks, influence on political discourse or capacity to set media agenda — it happened on Twitter.

Dataset

Historically, 2 major parties involve in Indian politics but in last election, there was another party AamAadmiParty, lead by Mr. Arvind Kejriwal, also came into the picture as third major party. In view of getting public opinion on candidates, 3 separate datasets, 1(One) for each candidate, has been considered. Election campaign period (Jan-Apr'15) has been taken into account to analysis the twitter data. Since, the focus is on classifying tweet data related to election candidate, only tweet publish time and content of the tweet has been selected to perform sentiment analysis to get desired outcome.

The download dataset of 100M Tweets on Indian Election from comp.nsu.edu site. was in json format which includes different attributes. The related columns were text, publish time, user name, user ID.

- Narendra Modi (1,500,866 tweet)
- Rahul Gandhi (409,146 tweets)
- Arvind Kejriwal (837,115 tweets)

http://www.comp.nus.edu.sg/~tankl/cs5344/data/India_Election.zip

Problem Definition

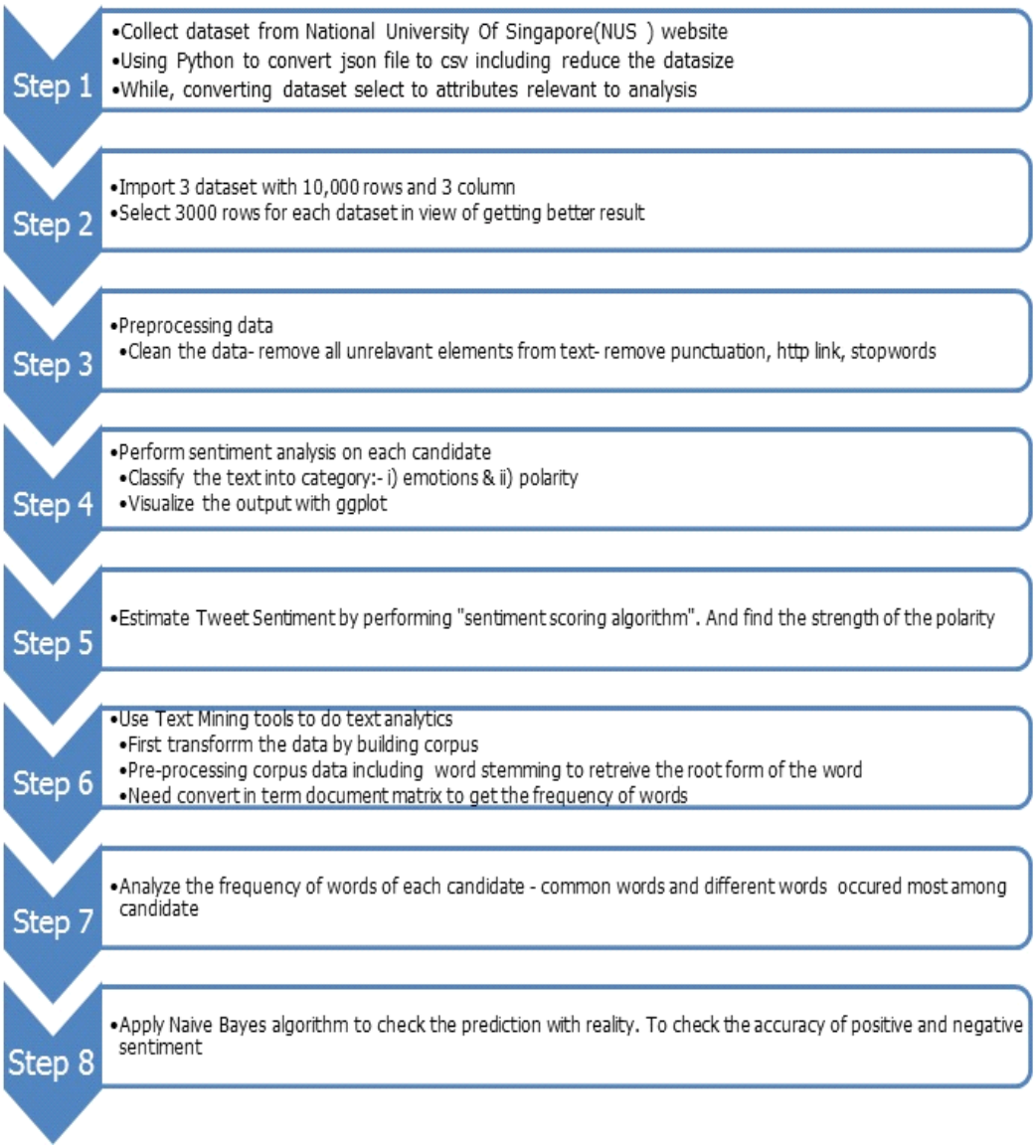
Candidate face challenges to get honest public opinion and expectation from mass people during "India General Election 2015" campaign 2015 from every corner of the nation. Which eventually helps candidate to design effective campaign plan to win the election.

Research Goal

Indian twitter user has had greater impact on this huge election, sentiment analysis has been considered here to analyze all the tweets related to election to understand public opinion on candidates.

Approach

In view of analyzing the data, following are the steps need to perform:



Step 1: Collect data

Collect tweets from twitter data related to 3 select candidates. Only content of the tweet and tweet publish timeframe have been considered here as the focus on public opinion towards candidates. Python programming language has been used to convert text file to csv file and reduce the size of the raw dataset before importing to R environment. Python is a dynamic, and multipurpose programming language. So, it is easy to handle large data in Python. With the help of Python, 2(two) attributes with 10,000 data has selected from each of 3 datasets.

Step 2: Import dataset into R environment

After loading 3 datasets, generate 3000 random sample data of each candidate to get better result out of the dataset.

Step 3: Preprocess the data

Since this is an unstructured data, clean the data is required before using this text data. Cleaning data includes removing punctuation, number, stopwords, http link.

Step 4: Perform Sentiment Analysis on each candidate

Classify the text based on emotions categories and polarity return with higher count. Learning method has been used here to classify the text. Use ggplot function to visualize the output of each candidate. Eventually sentiment analysis speaks the overall contextual polarity of a document.

Step 5: Estimate Tweets sentiment

Get the score of negative and positive sentiment by performing "Sentiment scoring algorithm". Here to get the result, "Breen's Approach" has been considered. This method is called Lexicon which is nothing but word dictionary

Step 6: Text Mining

At this stage, introduce "Text Mining" machine learning tools to determine the characteristics of the message. First, transforming text by building a corpus, that is a collection of text documents. Pre-processing text includes removing stop-words, changing letters to lower case, removing punctuations and numbers. After that, stemming words t retrieves the root from, so that words look normal. Term Document Matrix function provides the frequency of terms in document and also importance of the words by using performing find-Assoc. function.

Step 7: Analyze the frequency of words of each candidate

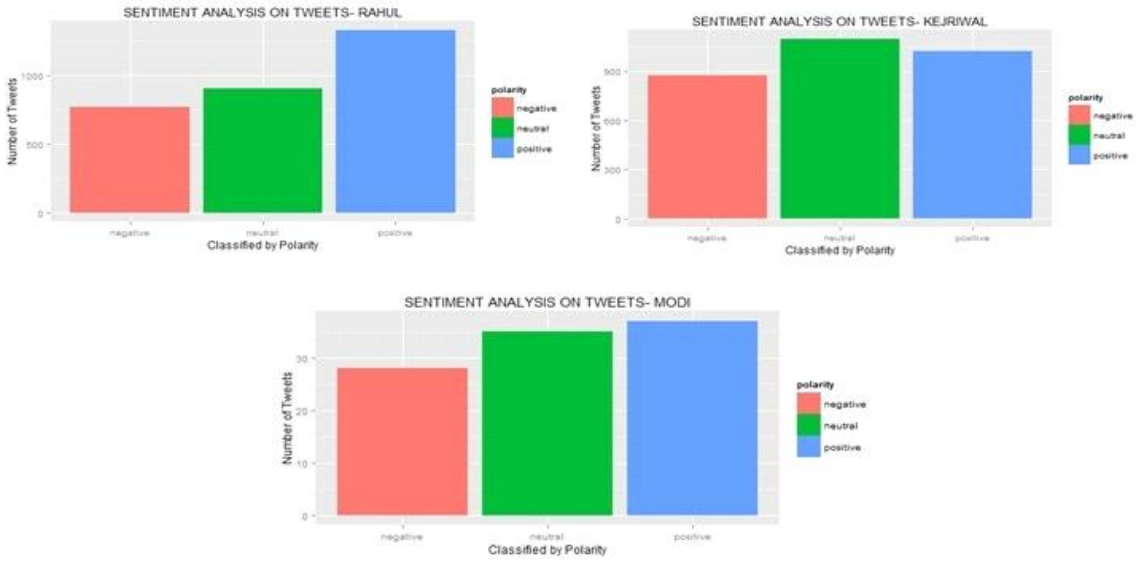
Comparison cloud and commonality cloud function have been using here to find out the words of different frequencies and the words in common from the text referring to public opinion on candidates.

Step 8: Check the accuracy of the model

Finally, text has been classified as either negative or positive by using Naive Bayes algorithm. Task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

Result

Sentiment analysis by polarity shows what people think about candidate



Among 3 candidates Rahul and Modi has greater positive sentiment than Kejriwal on twitter. So, both Rahul and Modi are likely to get more votes than Kejriwal. On the other hand, Kejriwal has more neutral sentiment compare to other 2 candidates. It may depict that people may not sure about Kejriwal election manifesto.

Visualize top occurring terms in term-document matrix for tweets on 3 candidates

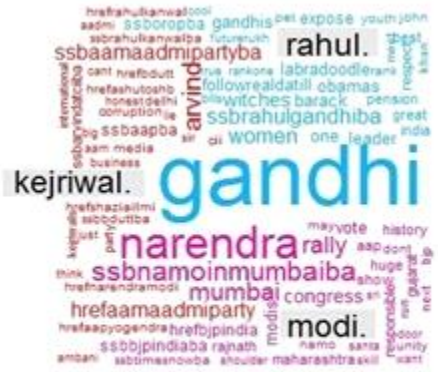
"Women" and "India" appear more often than any other word in most tweets related to candidate Rahul. Most likely he has been trying to talk about engaging women more for the development of the country.

Aam Aadmi Party (AAP) formally launched its political activity on 26 November 2012. And soon gained popularity in the country. Word "Aamaadmiparty" appear most due to party's self-promotion and highlighting its party manifesto. People talk about party manifesto and their future direction.

Whereas, Modi's BJP party hold rallies one after another during campaign period. As a result, Modi's predominance appears more clearly by comparing the occurrence of his name including "ralli" term in election-related tweets versus other names and relevant search terms.

Visualize words of different frequencies and words in common from all the tweet referring to public opinion

Commonality Cloud



Comparison Cloud

Commonality cloud shows that people talk about more on "India" and mass "People", when it comes to 3 candidates. It means mass people likely to expect that future leader would work for betterment of the common people of India. Whereas, comparison cloud show "gandhi" word appear most while talk about the Indian candidates. People likely to see Gandhi and Nandra Modi as there were the leading contender to take charge of the country

Apply Naive Bayes Classification model

Naive Bayes Classifier for tweet sentiment on 3 Candidates

```
          Reference
Prediction negative neutral positive
  negative       17       0        0
  neutral         0       7        0
  positive        0       0      576

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9939, 1)
    No Information Rate : 0.96
    P-Value [Acc > NIR] : 2.305e-11
```

Rahul

```
          Reference
Prediction negative neutral positive
  negative       35       0        0
  neutral         0     475        0
  positive        0       0       90

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9939, 1)
    No Information Rate : 0.7917
    P-Value [Acc > NIR] : < 2.2e-16
```

Kejriwal

```
          Reference
Prediction negative neutral positive
  negative        0       0        0
  neutral         0      14        0
  positive        0       0        6

Overall Statistics

               Accuracy : 1
                 95% CI : (0.8316, 1)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.0007979
```

Modi

Here, Naive Bayes model compare the prediction with reality. Result shows that, it is a very good model as the accuracy rate is 100%. Which means people predict negative and positive sentiment accurately for all the 3 candidates.

Conclusions

Analysis shows that people mostly talk about positive contribution about Rahul and Modi. They are 2 candidates' people most likely wants to form a new government. Although, Kejriwal enter into politics with mass popularity, during campaign period, Kejriwal face tough challenge to establish their party manifesto, as a result he gets more tweets which classified by neutral.

Finally, the analysis shows that how politicians and citizens cooperate to create an e-democracy based on information dissemination and political awareness especially during elections. Further to add that more successful parties used Twitter to push timely updates on online and offline campaign activities, to their followers.

https://github.com/Syedtanzin/Sentiment-Analysis--Capston-Project.git