PoliTweets – Predicting the Results of an Election using Sentiment Analysis

Ravi Patel
(001318953)
rpatel4@albany.edu
Department of Computer Science

Anand Patel
(001324098)
apatel4@albany.edu
Department of Computer Science

Saurabh Kamath
(001313200)
skamath@albany.edu
Department of Computer Science

1. ABSTRACT

These days, social medial is one of the major platforms for expressing emotion or feelings by public user for specific topics or subjects. One such social media site is Twitter where there are millions of users. A large portion of the researchers dealing with utilizing Twitter to monitor people reactions in political activities like debates and campaigns. Similarly, we use twitter data to find out the winner out of two political parties. We consider one ongoing election of state of Karnataka in India. We have used tweeter rest API for gathering the data then we have make the sentiment analysis of that data and with the help of association rule mining we have generated result for election.

KEYWORDS

- Social Media
- SVM
- Election Prediction
- Twitter
- Karnataka
- BJP
- CNG
- Sentiment Analysis

2. INTRODUCTION

In today's world, social media has major impact on people's life. Social media has changed the nature of information in terms of availability, importance and volume. An Election is the most important part of democracy. It is the instrument of democracy wherever the voters communicate with the representatives. There are around 336 million people uses twitter. So, if we use such large dataset from twitter for predicting election result, it will give better result. Twitter is a social media service that allows users to post "tweets" to the site [9]. These can either be viewed publicly by anyone who wishes to see or can be made private so that only people who have been allowed to follow a user can see that person's tweets. Regardless of the privacy, one of Twitter's features is that each tweet is limited to 140 characters, which includes whitespace characters, non-ASCII text, or links to web pages or images.

We are predicting the result of ongoing election held in Karnataka state, which is in India, the election is on mid-May 2018. We have selected main two parties(sides) for our prediction. Which are basically 'BJP' and 'Congress'. The reason behind selecting this two main or most influenced parties is both are always being the good candidate for winning the election in state. If we look to the history [1], we can see that before 2000 Congress rules over Karnataka but after 2000 BJP wins all the election. So, we are assuming that this year as well one of this two parties will win.

3. RELATED WORK

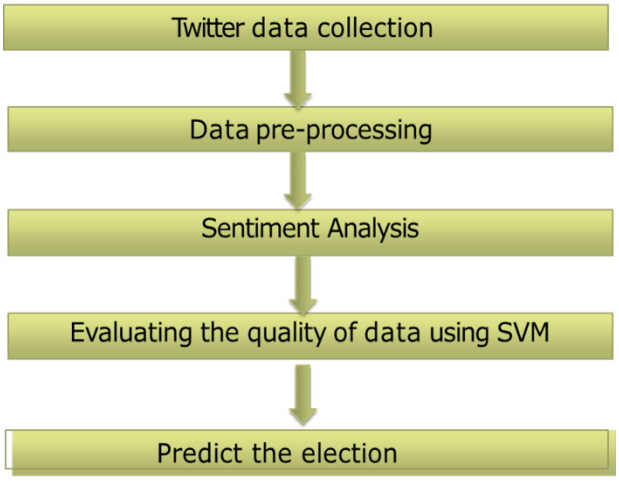
In this section, we are going to show some related works about predicting the result of an election using Twitter. Researchers use a different approach, there are researchers who try to discover the political preference of a user, then relate it to the election and there are others who use selected tweet related to the upcoming election and figure out vote preference of the user using that data.

The first method is by using selected data for some few days. The result can be observed by comparing number of tweets for each part. In this, they also consider number of negative and positive tweets for each side.

Researchers have tried to compare these two methods, for example, that tried to predict congress and senate election in several states of the US. They showed that both the method is the same, the prediction error can vary greatly. The research also showed that lexicon-based sentiment analysis improves the prediction result, but the improvements also vary in different states. Same result was shown in where they predict the result of Irish general election using both methods and which predicts the Italian primary election. All the research showed that sentiment detection does reduce the error of the prediction result. Because of that, several researchers focused on improving the sentiment analysis, such as and who used more sophisticated sentiment analysis than lexicon based in the US presidential election, France legislative election, and Italy primary election.

4. PROPOSED APPROACH

The following model we used to predict our result.



We used five steps to carry out prediction model. In first step, gathering data, we have used TWEEPY API to collect data from twitter. So, from there we can get good amount of information which can be helpful for analysing data and used that data for future use for example creating recommendation system.

The data we collected was very noisy, it is necessary to pre-process the data. We have removed unnecessary data like stop word, links, hashtags, tweet ID's, references tag, time, blank space, etc.

Then we have done sentiment analysis of the data and divided the data into positive and negative data based on the text which we got from tweeter data collection.

In next step, we have evaluated the quality of data using support vector machine (SVM) from divided data into training dataset and testing dataset.

In the last phase, we have generated a prediction model for predicting winner of Karnataka election by using modified polarity lexicon equation.

5. SYSTEM DESIGN AND IMPLEMENTATION

5.1 ARCHITECTURE

Election being held in most of the place in world. Sometimes it is helpful as well as interesting to predict the outcome of election before election on the basis of public interest. For that Twitter is one of the good platform to find out the winner of election. For getting result from twitter we have defined following prototype for our model.

- Gathering data (Used TWEEOY API)
- Data Pre-processing (Removed stop words)
- Sentiment Analysis (Positive or Negative)
- Evaluate the quality of data (Used SVM)
- Predict the result of election (Modified polarity Lexicon)

5.2 Dataset

Tweepy API is used to read tweeter data and as well as for store that data. Using this API we have collected a data about election for Karnataka state of India.

For collecting data, we have used different keywords for each party based on trend. (Congress and BJP).

For Congress we collected 5000 tweets using following keywords:

"karnatakElection2018, karnatakaElection, karnatakElection, Congress, Cong, rahul, RGinKarnatak, karnatakElection2018, ExitPollKarnataka, ExitPoll, karnatakElections2018, CongressMuktBharat, pappu, CongreaaGayi, congresscheatsdemocracy, CongressExposed, congressfails, congresschor"

For BJP also, we have collected 5000 tweets using following keywords:

"karnatakElection2018, karnatakaElection, karnatakElection, BJP, Modi, NarendraModi, modiinkarnatak, karnatakElection2018, karnatakaTrustModi, ExitPollKarnataka, ExitPoll, karnatakElections2018, BJPFAILS, BjpExposed"

For both parties we used geo location of Karnataka state. We used radius of 200mi and values for latitude is 15.3173 and for longitude is 75.7139.

5.3 DATA PREPROCESSING

There were some redundant data in our collected data. which should be removed for accurate result. From collected data we have lots of irrelevant information which needs to remove to make our dataset small and efficient for our model. So, for that we removed:

- Time of tweet
- Source
- Language of tweet
- Username
- Tweet id
- Profile image etc.
- Background image

So, after removing all this not needed information we came up with only text required for our model. Still our tweets contain some unneeded data which are listed below:

- Stop words
- Emotions
- Slang words
- Hashtags

After removing all above listed things from above dataset and by analyzing data, we came to know that there are some data which are not related to our project. Like there were plenty of tweets which was about US Congress and which was not useful for our data, so we find some keywords, and we removed those tweets, which contains that keyword. We also used tool, textcrawler for remove retweets.

After all process in both dataset. We came up with tweets, 5000 for BJP and 5000 for Congress.

5.4 GETTING TRAINING AND TESTING DATA

After clearing out data, we need to divide our data into two files one is training and another one is testing. In order to make model successful, set use for training and testing data must be good.

Training Data:
It is a collection of dataset which is used for training the classifier. The quality of data in this set should be reliable because it makes a huge impact of training data.

Testing Data:
It is a collection of data set which is used for discovering the predictive relationship between data. It is used for accessing the strength and utility of predictive relationships between data.

For our model we used around 4500 tweets as testing data and around 900 tweets as training data for each Congress and BJP.

5.5 SENTIMENT ANALYSIS OF TWITTER TEXT

To determine which tweets are positive or negative for each party, we need to do Sentiment Analysis for each tweet. The reason to do this is to understand the view point of each user for party and the reason behind posting that tweet. Those who have more positive

tweets are consider as winner of that election. Example of sentiment analysis is as follow:

“The Congress and Congress culture bring with it these evils. It is important to keep Karnataka away from Congress.”

The above tweet is a sample from our dataset, this tweet is considered as negative for Congress party while if same tweet appears in BJP data set it will be considered as positive. We gave sentiment value 1 for positive tweet and the sentiment value 0 for negative tweet.

So, for above example the sentiment value for BJP is 1, and for Congress the Sentiment value is 0.

To perform this functionality, we have used SVM classifier.

5.6 FREQUENCY OF WORDS

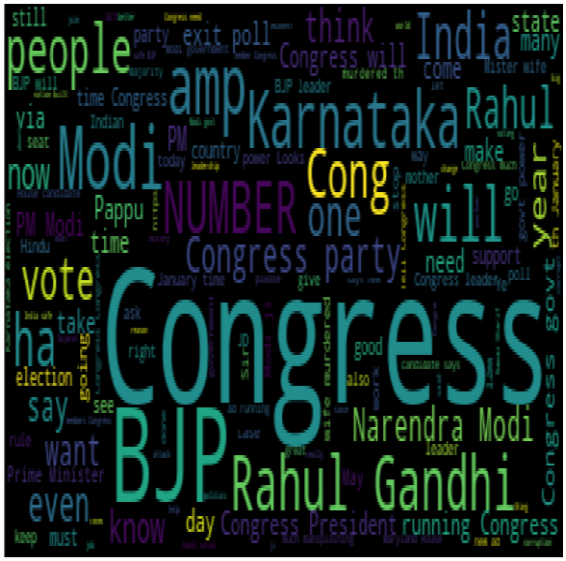
From our dataset, we have generated a word cloud, which is helpful to finding popular words for each party. This type of analysis is helpful to find the most frequent word for that party.

The word cloud image for BJP:

10 features from our data:

- president,
- karnataka
- congress
- modi
- cong
- india
- wrong
- bjp
- party
- Rahul

For Congress:

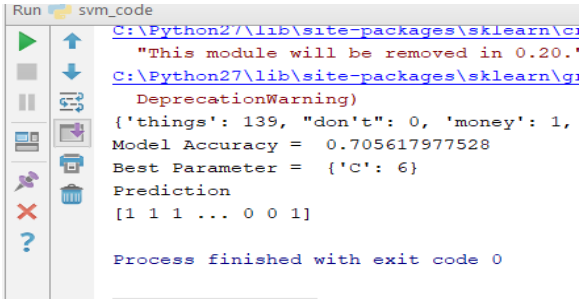


In this dataset, we are getting the same words but the frequency for Congress related words are increases.

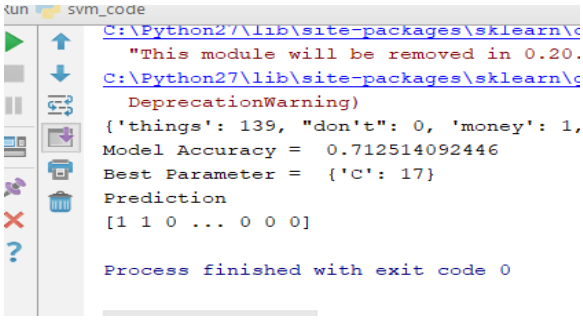
5.7 EVALUATING THE QUALITY OF DATA

To identify quality of data we have used SVM model. The below image shows the accuracy of our dataset.

BJP:



Congress:



6. RESULT AND ANALYSIS

To get the final result of election, we used Polarity Lexicon model modified by GayoAvello.

Equation used as shown below:

Equation 1: Modified Polarity Lexicon

$$(c_1) = \frac{pos(c_1) + neg(c_2)}{pos(c_1) + neg(c_1) + pos(c_2) + neg(c_2)}$$

Table 1 gives a description of Equation 1:

Description	
c_1	Candidate 1
c_2	Candidate 2
$pos(c_1)$	Positive tweets for candidate 1
$pos(c_2)$	Positive tweets for candidate 2
$neg(c_1)$	Negative tweets for candidate 1
$neg(c_2)$	Negative tweetss for candidate 2

Table 1: Equation Description

Below are the statistics from our dataset. Which contains the number of positive and negative tweets.

Party	Total Tweets	Negative	Positive
CONGRESS	**4434**	**3288**	**1146**
BJP	**4416**	**1118**	**3298**

After obtaining all the data we calculated the probability for each party, here we set trade of parameter to 1.

For Congress:

$$c_1 = \frac{1146 + 1118}{1146+3288+3298+1118} = \frac{2264}{8850} = 0.256$$

For BJP:

$$c_2 = \frac{3298 + 3288}{1146+3288+3298+1118} = \frac{6586}{8850} = 0.744$$

Here from calculation, we can see that BJP has higher probabilty than the Congress, it shows that BJP will win the Karnataka Election 2018.

7. LIMITATION AND CHALLENGES

The first problem we faced was getting tweets from API. We were getting some tweets in different language like Hindi, Urdu etc. So, this make my tweets unformatted. Then we took tweets which were only in English, which solved our problem. To preprocess data, we tried to write code by ourselves, but it doesn't work then after we found out library which solved our problem.

8. REFERENCES

[1]https://en.wikipedia.org/wiki/Elections_in_Karnataka

[2]https://www.digitalvolcano.co.uk/textcrawler.html

[3]Cameron, M. P. (2013). Can Social Media Predict Election Results? Evidence from New Zealand. No. 13/08

[4]Pak, A. &. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC.

[5]Gaurav, M. S. (2013). Leveraging candidate popularity on Twitter to predict election outcome.Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM., 7

[6]A Review: Prediction of Election Using Twitter Sentiment Analysis by Pritee Salunkhe, Avinash Surnar, Sunil Sonawane.

[7]Data Mining for Social Media Analysis:Using Twitter to Predict the 2016 US Presidential Election by Kabir Ismail Umar, Fatima Chiroma.

[8]Twitter Based Election Prediction and Analysis by Pritee Salunkhe, Sachin Deshmukh

[9]Makazhanov, A. R. (2014). Predicting political preference of Twitter users. Social Network Analysis and Mining, 1-15.

[10]Spoelstra, L. S. (2012). Prediction US Primary Elections With Twitter.

[11]Shung, M. C. (2012). A sentiment analysis of Singapore Presidential Election 2011 Using Twitter data with census Correction. problem.