POLITICAL MOOD DETERMINATION USING TWITTER DATA BY SENTIMENT ANALYSIS

Ruksar Begum[1], Varsha S[2], Nagma Neha[3], Aditi Ravichandra[4]

[1,2,3] UG Student, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India
[4]Assistant Professor, Dept. of Computer Science and Engineering, Atria Institute of Technology, Karnataka, India

---***---

Abstract - *Social media has changed the way users interact with each other, it is used as a platform to express polarized opinions on a global or specific context. Twitter is one of the powerful social networking site that generates millions of tweets per day and provides us live access to opinions. Twitter users discuss on subject matters such as government, technology, entertainment, politics, economy and so on. There is ample of data available but it is not organized according to the subject. On that account, it is required to classify the data so that some conclusion can be drawn out of it. This project develops a software that applies mood analysis methodology on tweets to anticipate social sentiment on political parties. The software takes input as the aspect for which user is interested to know the public's view with reference to the party. The factors on which parties are judged could be the parties' contribution towards the field of education, infrastructure, schemes, agriculture and many more. These factors can be dynamically updated based on real-time scenarios like political disputes, social issues, etc. The output graph displays summarized feedback for factor wise polarity for individual party, comparison of parties on individual factors, comparison of all factors. The parties are ranked across the factors.*

***Key Words*: Sentiment Analysis, Naive Bayes, Twitter, Machine Learning, Natural Language Processing, Mood Detection**

1. INTRODUCTION

The phenomena of Online Social Networks (OSN) has a global impact among web users. Social sites aim to create, share and express information on a real-time basis that can be useful to analyze in terms of data streaming and opinion diffusion [1]. The users are interested to know public perspective on the political parties. To collect diverse dataset of current opinion, twitter is used as a social platform wherein the users can post in short messages called tweets with a maximum length of 140 characters that allows to interact with celebrities, politicians, opinion leaders and other users. The project aims to solve the practical problem of understanding the current opinion towards the parties through public's live tweets rather than small localized polls organized by mainstream media corporations. To determine political mood, each tweet is analyzed and essential information is extracted by applying sentiment analysis technique. For each party, the sentiment score is tallied and ranked based on the factors.

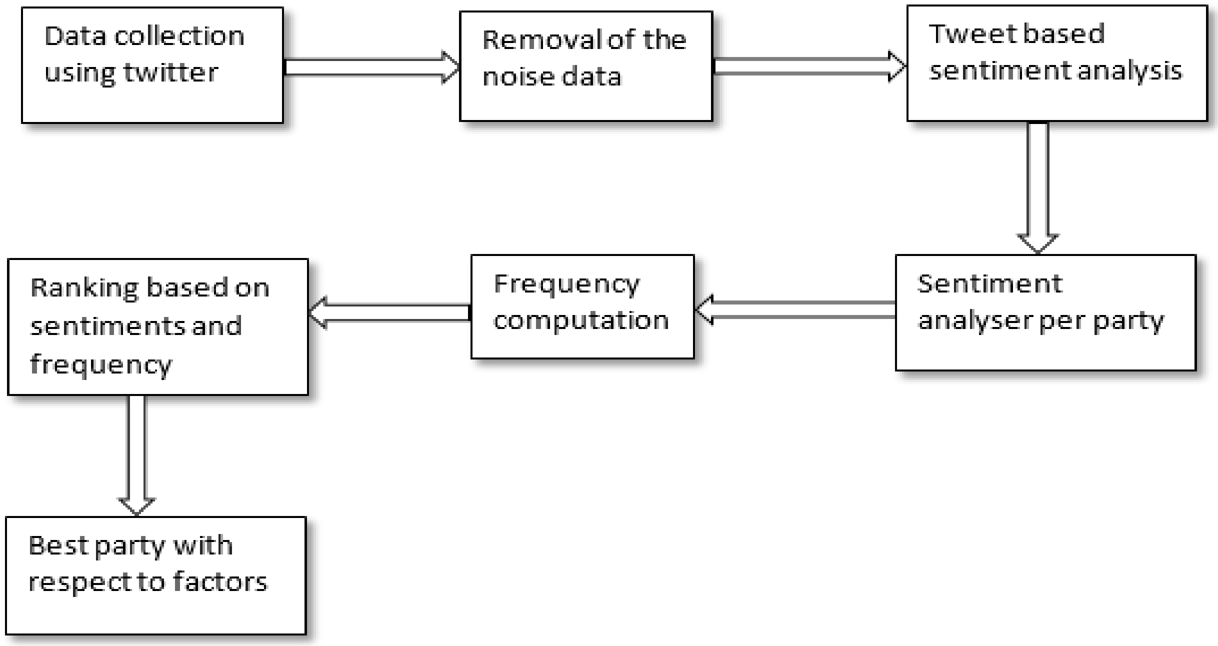
2. LITERATURE SURVEY

Paper 1 gathered tweets for candidates -Donald J. Trump and Hillary Clinton during the presidential elections of United States in 2016. The tweets where pre-classified with positive and negative labels by the naive bayes classifier. One of the major step is to weight extracted features so that significant words could be selected in order to increase the classification accuracy. Once the classification was completed, positive and negative tweets where scored on a daily basis as input for the prediction task. The objective of predicting was to find a correlation between user's mood polarity and a trend which clearly does not establish a threshold for predicting which candidate is going to win elections but can be useful to observe online behavior towards political issues [1].

In paper 2, tweet text is correlated to emoticons and related feelings about the candidates are categorized by creating mapping of emoticons. Maps where constructed for sentiments of tweets, metadata was used to note the location of the users. Twitter API does not have entry for the latitude and longitude co-ordinates. Hence, user inputted the location. Hand labelling of each location was required to determine state in which the user lived. Using the hand labelled data, a heat map for each candidate was created that shows the concentration of tweets from each state in comparison to the total number of tweets of the candidate. Lastly, for the states with at least one tweet they added the emoticon representing the dominant emotion of the tweets from the state [2].

3. PROPOSED MODEL

The vast data is evaluated to extract necessary information. Along with extraction, data cleaning, data integration, data transformation, data's sentiment score, frequency computation and data presentation steps are performed.


Data collection using twitter
Removal of the noise data
Tweet based sentiment analysis
Sentiment analyser per party
Frequency computation
Ranking based on sentiments and frequency
Best party with respect to factors

A) Data Collection:

Data from twitter can be collected only after the authentication process is successful. The authentication process makes use of OAuth Authentication API in which secret key and token is provided. After successful authentication, developers are allowed to gather corpus of tweets for hashtags. The hashtags for which tweets are required are specified and these tweets are stored in tweet storage. The data about the tweets is stored terms of TwitterId, TwitterDesc, UserId, Hashtag, Party. Tweets can be collected every time user wants to know the current opinion as the public's perspective fluctuates quickly due to the interviews, debates, responses to events and other issues that arise.

B) Tweets Pre-processing and Cleaning:

To avoid working with noisy and inconsistent data, pre-processing of data is an essential step where in the raw text is made ready for mining. By doing so it becomes easier to extract information from the text obtained from the tweet and apply machine learning algorithms on it.

Tokenization: In the process of tokenization, strings in the tweets are split into list of words, phrases or other meaningful elements termed as tokens. The tokens may be words, punctuation marks, numbers or hyperlinks.

Noise removal: Tokens that are frequently occurring and do not contribute to the analytic result are considered noise. They do not add any deeper meaning to the phrase, hence should be removed from the list of tokens. Noise could be twitter handles (@username), URL's, punctuations, numbers, special characters or stop words. Lexicons like emoticons are considered candidate textual features because they represent textual emphasis on tweets [1]. Noise is removed from the corpus C and normalization step is performed to convert into lower case. The new corpus C` is noise free.

Stemming and Lemmatization: Stemming is the process of reducing derived or inflected word to its stem or base word by removing the suffixes. For example-run, running, runs are all different forms of the word run. The process of stemming works only on single word without understanding the intended meaning but only considers grammar rules. Whereas, lemmatization groups different forms of inflected words so that they can be analysed as single item. It correctly identifies meaning of word in the sentence and its intended part of speech. Stemming and lemmatization is performed on corpus C`.

C) Classification:

Detecting and classifying sentiment polarity on set of tweets is achieved by training particular inputs with positive or negative mood tags. We perform the classification task using Naive Bayes Classifier [1].

D) Sentiment Analysis per tweet per feature:

Step 1: Obtain the tweets for required hashtags from tweet collection

Step 2: Count the number of tweets and assign it to N $_{\text{tweets}}$

Step 3: Iterate from first tweet till N $_{\text{tweets}}$

i) Each tweet is tokenized.
ii) For the list of tokens, positive, negative and neutral sentiments are computed for each factor.
iii) For every feature, sentiments throughout the tweet are added.
iv) Sentiments are stored in this format.

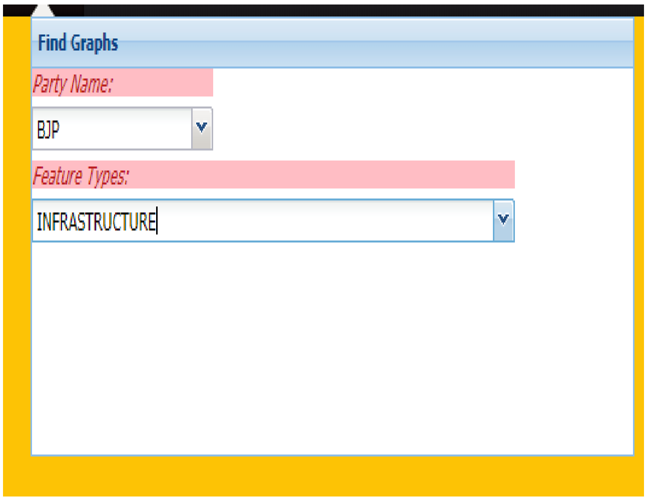
Tweet ID	Feature	Party ID	Positive sentiments	Negative sentiments	Neutral sentiments

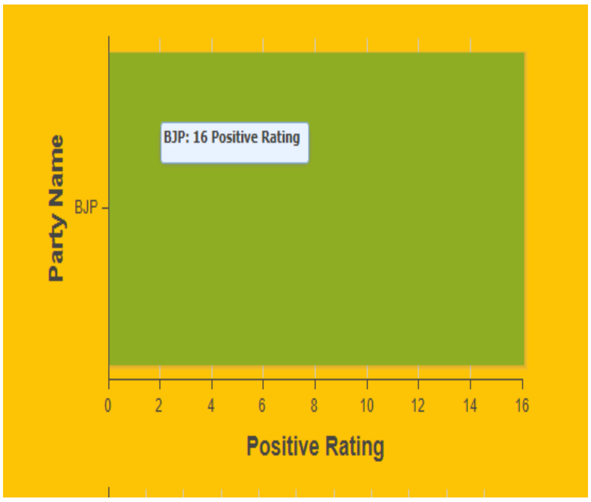
Polarity is computed per tweet per factor and a database is created that stores positive rating, negative rating, neutral rating and the party id.

E) Sentiment Analysis per party per feature
After tweet-based sentiment is obtained, maximum likelihood principle is applied to compute party-based sentiment and the final matrix obtained.

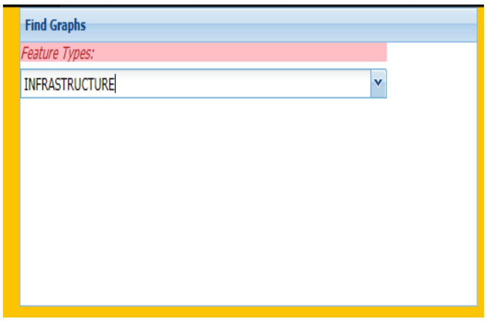
4. RESULTS

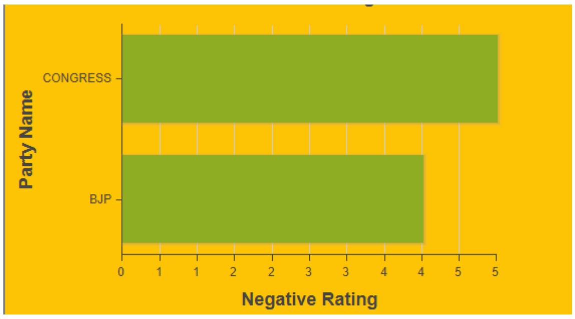
1. Feature wise graphs polarity:


Find Graphs
Party Name:
BJP
Feature Types:
INFRASTRUCTURE

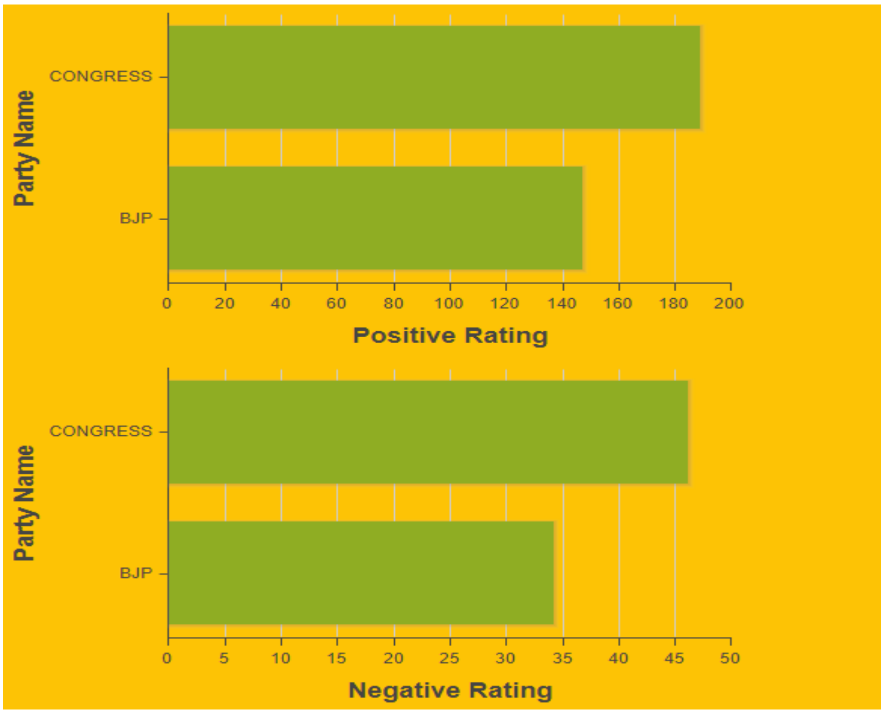


2. Feature wise comparison:


Find Graphs
Feature Types:
INFRASTRUCTURE



3. All features comparison:



5. CONCLUSIONS

- As the amount of information present on the internet is increasing, it can be used in the process of data analysis to give good conclusions.
- Using a combination of data aggregation techniques, natural language processing, linguistic analysis and popular visualization techniques we generated visually appealing and easy to understand graphs which provides summarized feedback.
- These graphs can be viewed to know public's perspective regarding parties and draw comparison between parties.

REFERENCES

[1] "Predicting Political Mood Tendencies based on Twitter Data" by A Hernandez-Suarez∗, G. Sanchez-Perez∗, V. Martinez-Hernandez∗, H. Perez-Meana∗, K. Toscano-Medina∗, M. Nakano∗ and V. Sanchez.

[2] "Analyzing Twitter Sentiment of the 2016 Presidential Candidates" by Delenn Chin, Anna Zappone, Jessica Zhao.

[3] K. Lee, D. Palsetia , R. Narayanan , Md. Mostofa A. Patwary , A. Agrawal , A. Choudhary, “Twitter Trending Topic Classification”, Proceedings of the 2011 IEEE 11th International Conference on DataMining Workshops, pp.251-258, 2011.

[4] S. Jain, V. Sharma, R. Kaushal, “PoliticAlly: Finding political friends on twitter”, 2015 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS), pp. 1-3, 2015.