Sentiment Analysis of Tweets for Indian LokSabha Election 2019 using IBM Watson NLU

Prakash C.O [1], Navneeth M.N [2], Nikhil D.B [2], Poojasree D [2], Archana J [2], Dr Shylaja S.S [3]

[1] Asst. Professor, Dept. Of Computer Science and Engineering, PES University, Bengaluru, India
[2] UG Student, Dept. Of Computer Science and Engineering, PES University, Bengaluru, India
[3] Professor and Head, Dept. Of Computer Science and Engineering, PES University, Bengaluru, India
coprakasha@pes.edu, navneethnarendra@gmail.com, ndb4263@gmail.com, pooja2000.dwarakanath@gmail.com, archanajayakumar4.aj@gmail.com

***Abstract* – These days, a huge number of user opinions are posted on every social media by thousands of users for any events, topic, current news, etc. Twitter is one of the biggest social media where people can post their tweets, share,and discuss it. Sentiment analysis is a study of people's feeling, reviews represented in textual form. The election is one of the most popular topic on social media. User's feeling or review on a related election is improving the national democratic process of elections. The proposed research work is machine learning-based tool for classifying the Indian Lok Sabha 2019 election-related tweets (Positive or Negative) which was collected for 3 months(Feb 2nd to Apr 25th) which spanned over 270 GB. Our research paper consists of two stages: Preprocessing, Sentiment and Emotion Analysis. We have discussed about the future prospects of this project**

Keywords – Twitter, Preprocessing, Sentiment Analysis, Indian Lok Sabha Election

I. Introduction

Social media has been a platform for people to express their opinion and views since many years, it is a place where people are free to speak freely without any restrictions and are completely allowed to share their opinion regarding a topic.
Can this be used to analyze future trends or events or maybe even predict elections ?
Research using data from social media and more importantly twitter has become quite prominent today. The community of the popular platform counts more than 350,000,000 active users at the moment. It has been considered as the best social media for data extraction. Although extensive research is being done in this area there are still certain challenges: like to find the optimal way of data gathering. Twitter provides a variety of API's for data gathering, Selecting the right API to gather most relevant information inorder to get accurate results.
This paper mainly focuses on analyzing data related to Indian Lok Sabha elections 2019. We have collected data related to various political parties and more specifically the two major national parties BJP (Bharatiya Janata Party) and INC (Indian National Congress). After the data has been gathered, it is filtered to gather only the required tweets it is then classified to different parties using keywords and finally applying sentiment and emotion analysis on the tweets collected.

The three main difficulties encountered in our task:

1. Data Gathering:

Inorder to collect data from twitter the first necessity is creating a developer account. Twitter offers various types of API's

i) Search API
ii) Streaming API

The search API allows the user to collect specific data of specific days, whereas the streaming API allows the user to live stream and gathers data based on the keywords given. The selection of the right API is very

important as the accuracy of the model depends on the purity of the data collected.

2. Pre-processing:

Twitter data is usually noisy, alot of unrelated tweets do get collected during the data gathering process, moreover we use only the text part of the tweet for our model for which we had to remove the URL's, Therefore pre-processing or noise removal is a very challenging task in NLP based projects.

3. Sentiment Analysis:

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral is called as sentiment analysis or opinion mining. There are differernt aprroaches to this problem which includes classifier based, using built in modules and dictionary based methods. Opinion mining can be used for different purposes like :

i) Buisness: Companies could improve their product after understanding the sentiment of the users given in their feedbacks.
ii) Politics: Sentiment Analysis can be used to understand the views of the citizens towards their leaders and the government and can also be used as feature in predicting the elections.
iii) Public Actions: Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

II. Background and Literature Review

Although Social media analysis started very recently alot of research has been worked on it inorder to find the view of the users. In our work we review seminal works that targeted elections or candidates that participated in elections

In 2016 Parnian, Alireza and Hamid [1] published a paper named "Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls" wjich focused on predicting the US Presidential elections. They had collected large data of more than 3,70,000 tweets, they had used a gaussian based regression

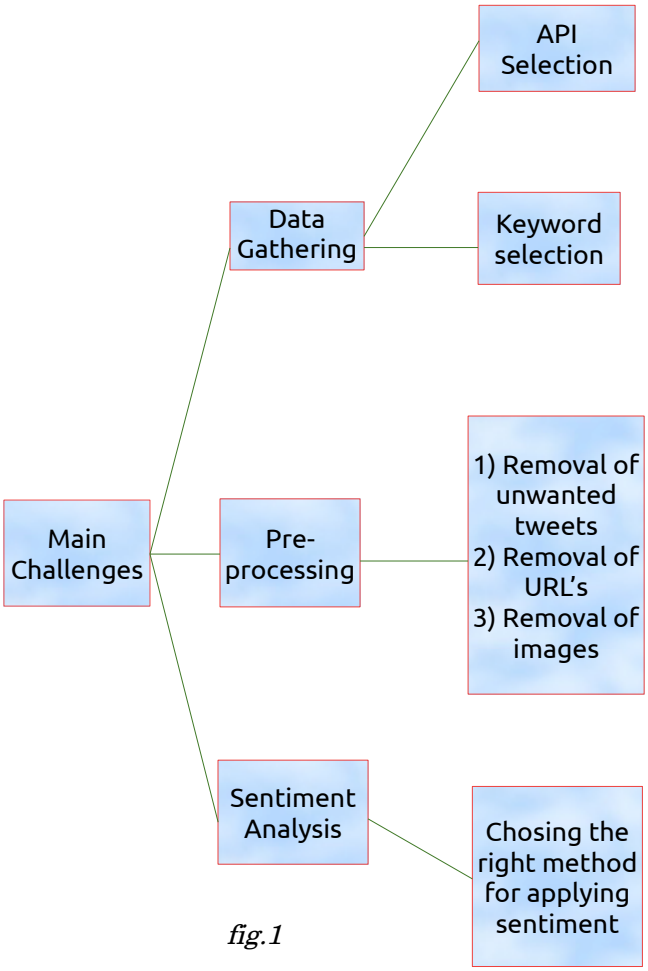


fig.1

model to the predict the elections. The model foresees election results at the beginning of the election week. Using the jackknifing (Efron, 1982) the error distribution of the model is estimated. The model was not capable of removing spam tweets which is very common in twitter.

In 2015 Joose and Chooralil [2] proposed a new method for finding the sentiment of tweets. They analyzed the tweets with lexicol resources such as SentiWordnet, Wordnet and word sense disambigution, they extracted knowledge and information from the tweets , this model showed the highest accuracy and

they also provided a negation handling approach in the preprocessing stage

Another interesting work by Akshi Kumar et al. [3] which was called “Emotion Analysis of twitter using Opinion Mining“. analysis. However, they used a slightly different approach. According to the authors the basic sentiments that a person might have are: Happiness, Anger, Sadness, Fear and Disgust. These are the sentiments that Paul Ekman and his team found in their research in 1972 to be the most common among human beings. The data that were gathered in this work were classified based on Ekman’s approach. This approach may find several different uses like recommendation systems in business intelligence and to reveal vote intentions of social media users.

Soler et al. [4], published a paper in 2012 on “Twitter as a tool for predicting election results”. This paper presented a new tool for Twitter analysis called TaraTweet. Using this tool nearly 5,00,000 tweets were analyzed and their sentiement was shown in the TaraTweet website. This paper showed sufficient accuracy, a few parties percentages predicted was different when compared to the actual percentages which is accpetable as the voter can change his mind even in the last minute. This paper indicates that Twitter analysis is a safe method to conduct experiments and come up with results that are really close to real-time vote intentions.

Andranik Tumasjan et al. contributed an amazing paper named “Predicting Election results with Twitter : What 140 characters reveal about Political Sentiment” [5] which focused on a number features for prediction. They collected nearly 1,00,000 tweets related to their 6 political parties and their famous politicians. To extract the sentiment of these tweets automatically, we used LIWC2007 (Linguistic Inquiry and Word Count; Pennebaker, Chung, and Ireland 2007), a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. They focus on 12 dimensions in order to profile political sentiment: Future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money. Following the methodology used by Yu, Kaufmann, and Diermeier (2008) we concatenated all tweets published over the relevant timeframe into one text sample to be evaluated by LIWC. Tweets were downloaded in German and automatically translated into English to be processed by the LIWC English dictionary. These are the 5 prominent research papers we have come across during our research. There are also othe successful models which have good accuracy. The works show how prominent this field is. In the later sections we will discuss how we came about our project.

III. Methodology

The proposed method comprises of three main features
1) Data gathering
2) Pre-processing
3) Tweets Classification
4) Sentiment Analysis

Data Gathering

We collected the Twitter data related to 2019 Lok Sabha Elections using the twitter API. Keywords like Modi, BJP, Congress, Chowkidaar, Rahul Gandhi, etc were given while live streaming. We collected 270 GB dataset from 2 nd February to 25 th April. It was collected in the .json format and had Unicode UTF-8. Understanding the features of each tweet in the dataset was quite challenging. The original tweets had the following features: created_at, id, id_str, text, source, truncated, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, user, geo, coordinates, place, contributers, is_quote_status, quote_count, reply_count, retweet_count, favorite_count, entities, favorited, retweeted, possibly_sensitive, filter_level, lang, timestamp_ms. The retweets have only one extra attribute - retweeted_status. It also has the details of the original person who tweeted it. If the truncated is false (i.e., if the length of the text exceeds 240 letters), then there is another feature called extended_tweet. In the .json file, the data is stored in the form of python dictionary data structure. The data was stored in different files week-wise. Hence, each file had millions of tweets and thus was difficult to work with. The processed dataset (.csv files) could not open due to maximum limit on the number of rows in a csv file. So, we split the week-wise .json files to day-wise .json

files. We used multi-threading for this. We used the json.loads() function from the json module to read the data from the .json file. We preprocessed this data and then stored it in a .csv file. To write it into the .csv file, we imported csv module. One of the major challenges we faced during preprocessing was that tweets related to the US Elections were also included in the dataset due to the keyword ‘Congress’ given during data collection. Spam tweets were also a part of the dataset. This made our data noisy. We’ve used some keywords to extract the tweets related to the Lok Sabha Elections 2019 and ignoring the rest.

Pre-processing:

In preprocessing , we converted all named and numeric character references like &gt, #62, in the text to corresponding unicode characters such as ‘\u003ca’ , ‘\u2026’, etc using the html.unescape() function from the html module. We replaced ‘&amp’ with ‘&’ and then with ‘and’. The daywise file size was huge. Due to live-streaming of the data, there tweets and retweets with the retweet count updated everytime it is repeated. Hence, we’ve taken only the latest retweet of the day, with the updated retweet count and the full text. This also reduced the file size. There were some words like ‘cooooool’ with the extra letters which we replaced by ‘cool’ or the original word using itertools. We also used the set_options() function from the preprocessor module to remove the urls, emojis, reserved characters like RT and FAV, and smileys like ;), :), xD, etc. We made use of the ASCII values to remove punctuations (like didn’t to didnt) and special characters (like‘c‘). We also converted the whole text into lowercase. We expanded the contractions like can’t to cannot. We appended only the required data to the .csv files (i.e., created at, text, retweet count, full text, tweeted by) after the preprocessing. The retweet count is needed to know how popular that particular tweet became which could be helpful to predict who would win. Each program took a lot of time to execute (almost a day) because of the large number of tweets around 10,000 to 8,00,000 per day-wise file. We also made a wordcloud of the non-dictionary words. A wordcloud is a visualisation of words typically associated with Internet keywords and text data. They are most commonly used to highlight popular or trending terms based on frequency of use and prominence. We used the Dict() function from the enchant module to check if the word is meaningful or not. We imported various modules like wordcloud, matplotlib.pyplot, collections and numpy for the same.

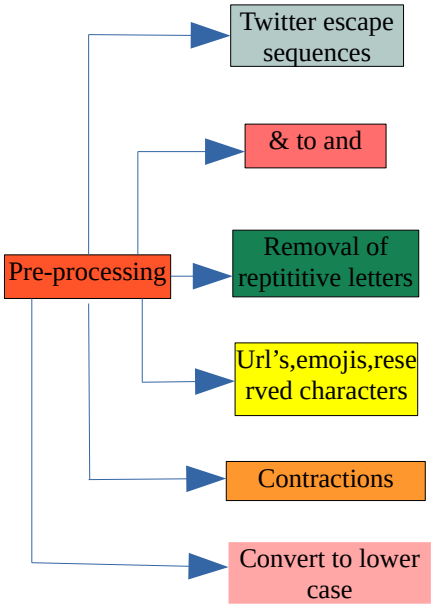


Fig.2

Tweets Classification:

As said earlier twitter data is very noisy. Alot of unrelated tweets do get collected as twitter is a free platform for users so there will be thousands of spam tweets. We had to first purify our data. Since we concentrate only on the text part for analysis only the text part of the tweet is taken into consideration. Other attributes of the tweet can also be taken which includes retweet count, geo-tagged location, hashtags, etc. There are various ways of removing the unrelated tweets, we chose keyword based removal. The tweets containing certain keywords were classified to their respective parties, for example if a tweet contains the word BJP or Narendra Modi or other politicians under that party, it would be grouped to a single category. So taking the tweets we want implies we are neglecting the tweets we dont want. Another challenge is to make sure the tweets talk about political issues related to only India. The geo tagged location of the person who has tweetwd was taken into consideration, since we only concentrated on the persons who would actually vote.

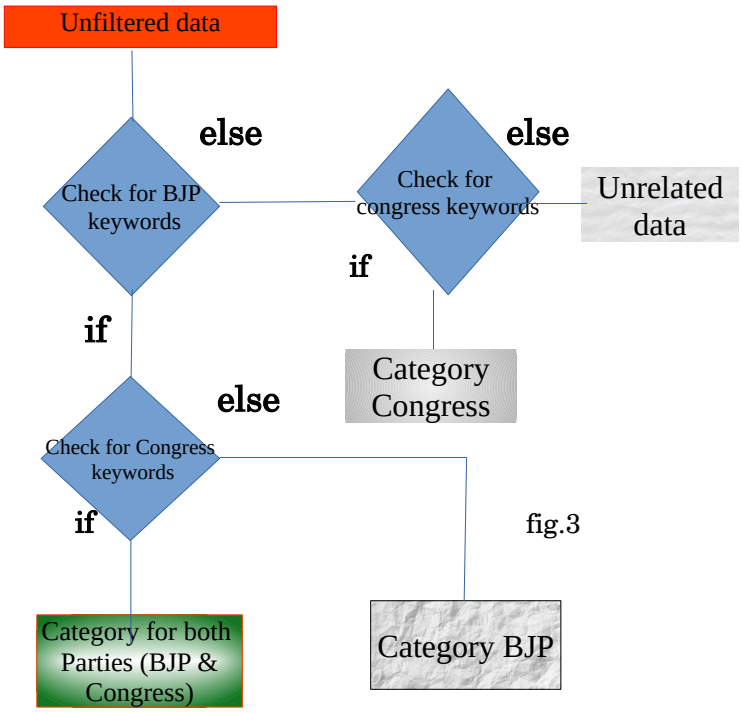


fig.3

Some of the tweets gathered are considered polarized beforehand. Tweets containing the hashtags “#bjp4india”, “#chowkidar” are the official hashtags in favor of Narendra Modi and more collectuvely BJP.

Sentiment Analysis:

It is clear that the main objective of this work is to clssify the sentiment of the tweets. The sentiment of the tweet justifies the the opinion the user has towards the specific topic. In our project the polarity of the tweet justifies the liking of the user towards that party. The tweets after being classified to different parties were tested for their sentiment. The sentiemnt label describes how inclined the user is towards that party. In our project we find out the day wise change in trends of the people towards a particular party. In 1972 Paul Ekman and his team had researced about the most common emotions in humans and concluded it to be Joy, Anger, Disgust, Sadness, Fear and Happiness. In our research we have found out the daily trends of change in these emeotions of the people towards Narendra Modi and Rahul Gandhi who were the most prominent politicians trending during the 2019 Indian Lok Sabha elections.

There are different methods for finding out the sentiment of a statement. In the below sections we will discuss briefly about a few of these these methods and their pros and cons.

1) Classifier based approach:

In this method the goal is to build a model which can be capable of labelling the tweets as ‘positive’ and ‘negative’ automatically. There are many classifier algorithms which can used for finding the sentiment like Naive Bayes classifier, Logistic Regression classifier, Lexicon classifier, Support Vector Machine (SVM) classifier, Random Forest classifier etc. Among all these Random Forest Classifier has been found to be the most accurate and efficient. For building a classifier model a predefined dataset should be used as training set . These models will not work if it has not been trained earlier. Therefore we have ensure the training set has wide variety of data. The more the variety the more is the accuracy of the model but certain measures must also be taken during training inorder to avoid overfitting and underfitting

2) Using python libraries like TextBlob , Vader Sentiement or SentiWordNet:

Certain prebuilt python libraries are available which can be used to find the sentiment of sentences. These are powerful tools which find the sentiment by subjective analysis. Subjectivity, is defined as the number of subjective words per sentence. A tweet that has a lot of words that express sentiment like adjectives, adverbs, verbs etc. has a higher subjectivity score than a tweet that does not. These tools are very effective for movie rewiews and product reviews but its efficiency drops down when it comes to political issues , since it has been trained with words in the English Dictionary and will not understand proper nouns. Therefore a tweet related to politics can have a number of proper nouns which includes politicians, parties, places. These tools fails to understand the context of the tweet which contains proper nouns due which it may not be that efficient in finding the sentiment of the tweet.

3) IBM Watson NLU:

IBM Watson NLU (Natural Language Understanding) is a tool provided by IBM for finding the sentiment of a sentence. We have used this tool to find the sentiment

of the tweets we had gathered. The efficiency of this tool has found out to be the highest among all other sentiment analysis methods. We first need to create an IBM cloud account to gain access to their services. IBM has provided a wide variety of services which include Visual Recognition, Natural Language Understanding, Natural Language Classifier, Tone Analyzer, Speech to Text and many more. Once the account is developed IBM offers various choices of API's . After obtaining the API our data was passed to find the sentiment of the tweets and to find the emotions related to Narendra Modi and Rahul Gandhi expressed by the people through their tweets. After obtaining the API key provided by IBM Watson we pass the tweets and extract information such as sentiment, keywords, entities and categories IBM Watson NLU gives the sentiment by a machine learning based approach. It has the capability to recognize proper nouns since it links the API calls with new articles to gather information regarding to the context and gives the sentiment with high accuracy. For example if a tweet contains the word Narendra Modi IBM watson recognizes its category as Bharatiya Janatha Party.

Proposed work Flow chart

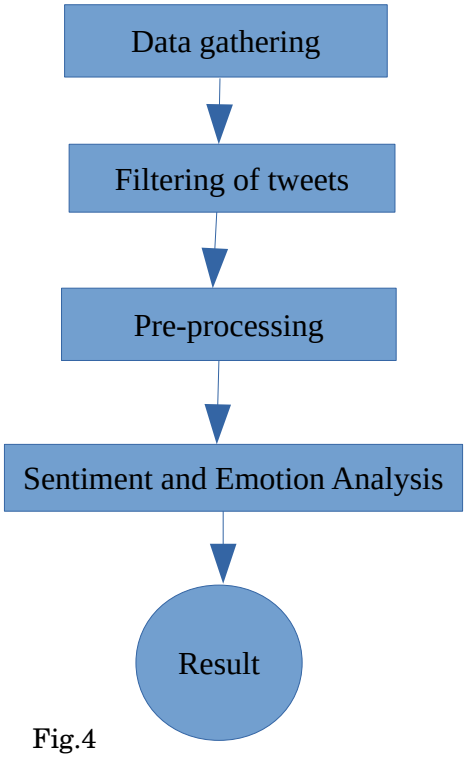


Fig.4

How IBM Watson works

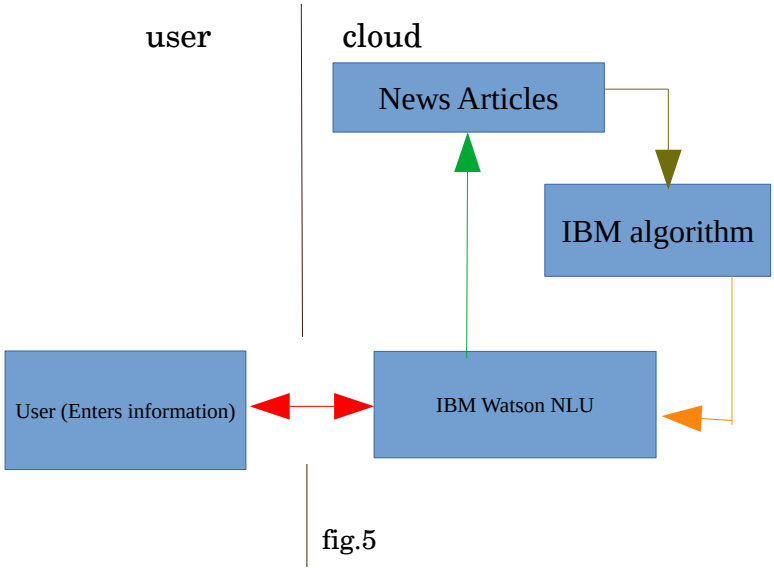


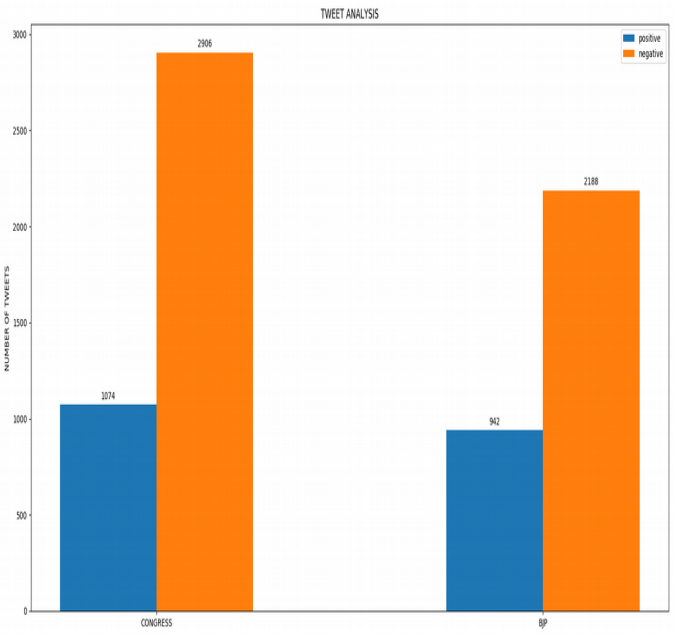
fig.5

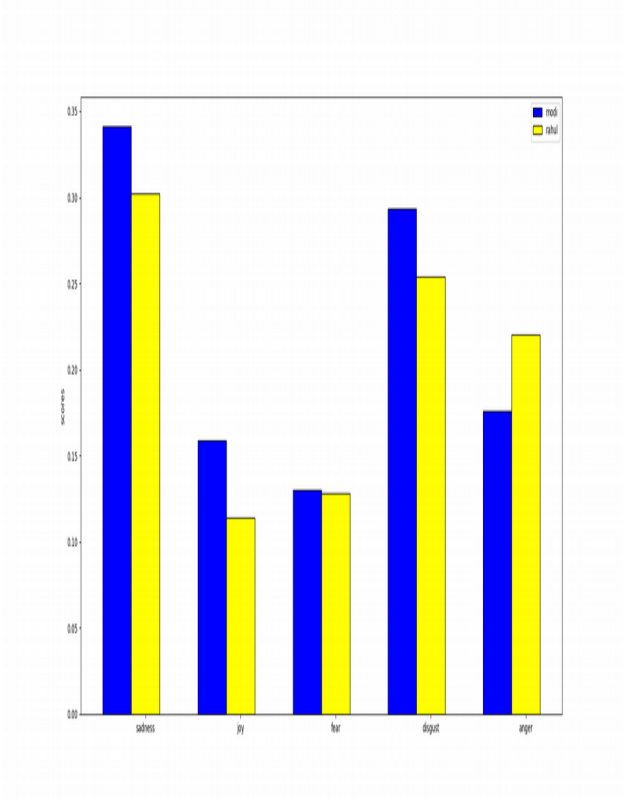
IV. Results and Conclusion

Each day tweets that were collected were passed through the IBM cloud and the results were plotted as shown for a single day

plot of positive and negative tweets towards BJP & Congress

Fig.6





Emotion plot (Narendra Modi vs Rahul Gandhi)

Fig.7

It can be seen than that the people have a more positive liking towards BJP than Congress through the plot. From emotion values for Narendra Modi and Rahul Gandhi we can infer that people have a liking towards Narendra Modi. Through these analysis we can fairly say that BJP has a higher chance of winning. We cannot predict the number of seats because prediction would require more features which we have not taken into consideration but this model can give an overall analysis of the trend of the people towards political issues. Therefore through this research we can conclude that sentiment analysis or opinion mining of tweets can give us valuable information regarding political issues which can be extensively used to predict events and help humanity in a much faster way. People have always wanted to know about the future therefore social media can be used a very rich resource to predict the future. Social media has been playing a major role in our lives today. It has made the entire world faster by connecting with people millions of miles away in fractions of seconds. We still have lots of oppurtunities to make use of this enormous data to help humanity in the most optimum way as possible

V. Future work

- More efficient algorithms can be developed for categorizing the tweets to different parites
- Sentiment analysis can be taken as a feature to build a model to predict any election using social media
- We have neglected tweets writen in local languages, only English tweets had been taken into consideration, the local language tweets can also be used by converting them to English

VI. References

[1] J. M. Soler, F. Cuartero and M. Roblizo, "Twitter as a Tool for Predicting Elections Results," Proc. IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 1194-1200.

[2] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation," Proc. Int'l Conf. on Control Communication & Computing India (ICCC), Trivandrum, 2015, pp. 638-641.

[3] A. Kumar, P. Dogra, and V. Dabas, "Emotion analysis of Twitter using opinion mining," Proc. 8th Int'l Conf. on Contemporary Computing (IC3) (IC3 '15). IEEE Computer Society, Washington, DC, USA, 2015, pp. 285- 290.

[4] T. K. Das, D. P. Acharjya and M. R. Patra, "Opinion mining about a product by analyzing public tweets in Twitter," Proc. Int'l Conf. on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-4.

[5] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Proc. 4th Int'l AAAI Conf. on Weblogs and Social Media, pp. 178-185, 2010.

[6] Kiran Sangada , Jitendrakumar Dhobi , "Sentiment Analysis of Tweets for Indian Election Using Random Forest Classifier"

[7] "TextBlob: Simplified Text Processing", https://textblob.readthedocs.io/en/dev/ [Accessed 2/2/2018].