International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 290-298

Twitter Sentiment Analysis for Election Prediction

V.RUPESH BABU
DR.ALTHAF ALI *(ASSISTANT PROFFESSOR)*
MASTER OF COMPUTER APPLICATION
MADANAPALLE INSTITUTE OF TECHNOLOGY&SCIENCE ANDHRAPRADHESH-517291

Abstract:
An election is conducted to view the public opinion, where a group of people choose the candidate by using votes, many methods are used to predict results. In the present day, social media platforms are playing a vital role in influencing people's sentiment in favour or against a government or an organization. We use Twitter data to analyse and predict the stand of future elections using sentiment analysis. Lexicon based approach with supervised machine learning and Naive Bayes algorithm is used to find the view point in tweets and predict sentiment score. The sentiment scores are then divided into positive, negative and neutral, to finally predict the winning party.

Keywords - *Election prediction, supervised learning, sentiment analysis, Naive Bayes algorithm*

Date of Submission: 25-05-2022 Date of acceptance: 05-06-2022

I. INTRODUCTION:

India is a federation which has a parliamentary system and is governed with the constitution of India guidelines. There is a central government which looks after the whole country and the states have their independent governments to look after the respective states.

India has the largest democracy in the world. It is a government formed by the people, of the people and for the people. One of the most efficient ways of government is a democracy if done in a free and fair way. The people elect their representatives to take decisions in the parliament and run the country. Citizens try to elect the candidate they think will bring some changes and help their community.

The general elections are held in a gap of 4 years where the citizens of India above the age of 18 are eligible to cast vote and choose the suitable candidate. The candidates are elected to be the voice of the citizens in the Lok Sabha whereas the Rajya Sabha candidates are indirectly elected. The election follows a hierarchical process where the people's representative elect the prime minister and president of India.

In less than 20 days, Indian voters from Kanyakumari to Kashmir will go to the polls to select their next parliament. The country's 2019 general election—like previous contests—will be the largest democratic exercise in world history. More than 850 million voters will be eligible to help determine which political party or alliance will form the government and, in turn, who will serve as prime minister.

With 29 states and different regional parties, it is difficult to decide which party will win the elections this time if we go by the old methods of manual queries to find out the sentiments and views of the people.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered. So what should a brand do to capture that low hanging fruit?

With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. We believe it is important to classify incoming customer conversation about a brand based on following lines:

1. Key aspects of a brand's product and service that customers care about.
2. Users' underlying intentions and reactions concerning those aspects.

These basic concepts when used in combination, become a very important tool for analyzing millions of brand conversations with human level accuracy. They are as follows:

Sentiment Analysis

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral. You can input a sentence of your choice and gauge the underlying sentiment.

Intent Analysis

Intent analysis steps up the game by analyzing the user's intention behind a message and identifying whether it relates an opinion, news, marketing, complaint, suggestion, appreciation orquery.

Contextual Semantic Search (CSS)

Now this is where things get really interesting. To derive actionable insights, it is important to understand what aspect of the brand is a user discussing about. For example: Amazon would want to segregate messages that related to: late deliveries, billing issues, promotion related queries, product reviews etc. On the other hand, Starbucks would want to classify messages based on whether they relate to staff behavior, new coffee flavors, hygiene feedback, online orders, store name and location etc. But how can one do that?

We introduce an intelligent smart search algorithm called Contextual Semantic Search (a.k.a. CSS). The way CSS works is that it takes thousands of messages and a concept (like Price) as input and filters all the messages that closely match with the given concept. The graphic shown below demonstrates how CSS represents a major improvement over existing methods used bythe industry.

A conventional approach for filtering all Price related messages is to do a keyword search on Price and other closely related words like (pricing, charge, $, paid). This method however is not very effective as it is almost impossible to think of all the relevant keywords and their variants that represent a particular concept. CSS on the other hand just takes the name of the concept (Price) as input and filters all the contextually similar even where the obvious variants of the concept keyword are not mentioned.

TWITTER USED FOR OPINION MINING

Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life everyday. Therefore microblogging web-sites are rich sources of data for opinion mining and sentiment analysis. Because

PROCESS OF SENTIMENT ANALYSIS

This is a general description of the process that we follow through the project to obtain thetweets and analyze them.

II. DATA COLLECTION

The data collection step is the initial phase in the project, where data is collected from twitter. There are two methods on how to connect and collect tweets from Twitter. The first method is by searching tweets matching to the keywords. The second method is by collecting all the tweets provided by Twitter through streaming API, or all the tweets in a specific language, or all the tweets in a specific location, then putting all of them into our database.

Both methods have their own advantages and disadvantages. For example, the first method requires only small storage as the data is relatively small. The downside is that we cannot get data from other keywords (if we need to) from an earlier time. Twitter allows the search API only for 7 days backwards. This data collection method is suitable if the focus is on the feature extraction or the prediction method. With the secondmethod, we can apply any set of keywords to get the best result.

As we are going for prediction and future analysis the twitter API extraction process is more helpful.

III. PREPROCESSING

Twitter analysis methods have various preprocessing steps of text. One of the most important goals of preprocessing is to enhance the quality of the data by removing noise. Another point is the reduction of the feature space size.

A. Lower Case Conversion: Because of the many ways people can write the same things down, character data can be difficult to process. String matching is another important criterion of feature selection. For accurate string matching, we areconverting our complete text into lower case.

B. Removing Punctuations and Removing Numbers: All punctuations, numbers also need to be removed from reviews to make data clean and neat. Unnecessary commas, question marks, other special symbols get removed in this case. Here, we are not removing the dot (.) symbol from our reviews because we are splitting our text into sentences.

C. Stemming: Stemming is the method of conflating the variant styles of a word into a standard illustration, the stem. For example, the words: "presentation", "presented", "presenting" could all be reduced to

a common representation "present". This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented. Stemming in our case is helpful in correct word matching and countingcase.

D. Striping White Spaces: In this preprocessing step all text data is cleaned off. Allunnecessary white spaces, tabs, newline character get removed from the text.

IV. SENTIMENT ANALYSIS

a) Machine Learning Approach

There are two approaches to machine learning, supervised and unsupervised. in our project we used a supervised machine learning approach. In a supervised machine learning approach, there is a finite set of classes for classification. Training dataset is also available. Most research papers do not use the neutral class, which makes the classification problem considerably easier, but it is possible to use a neutral class. Given the training data, the system classifies the document by using one of the common classification algorithms such as Support Vector Machine, Naïve Bayes etc. We used Naive Bayes algorithm for classification of tweets. We classified tweets into polarity and emotion also using Naive Bayes classifier. Naive Bayes is a machine learning algorithm for classification problems. It is based on Bayes' probability theorem. It is primarily used for text classification that involves high dimensional knowledge sets. A few examples are spam filtration, sentimental analysis, and classifying news articles. It is not only known for its simplicity, but also for its effectiveness. It is fast to build models and make predictions with the Naive Bayes algorithm.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where P(A|B): Probability (conditional probability) of occurrence of the event given the event B is true. P(A) and P(B): Probabilities of the occurrence of event A and B respectively. P(B|A): Probability of the occurrence of event B given the event A is true.

Naïve Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

b) Lexicon Based Approach:

There are three main approaches to compiling sentiment words. They are the manual approach, dictionary-based approach, and corpus-based approach. In our research, we used a dictionary-based approach. We used eleven different variables for classification, that variables are sadness, tentativeness, anxiety, work, anger, certainty, achievement, positive words, negative words, positive hashtag and negative hashtag. We collected various word related to those eleven variables and classified them microblogging has appeared relatively recently, there are a few research works that were devoted to this topic. In our paper, we focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, we build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that our proposed techniques are efficient and performs better than previously proposed methods. In our research, we worked with English, however, the proposed technique can be used with any other language.

Existing system:

The existing system works only of the dataset which is constrained to a particular topic.

The existing systems do not determine the measure to impact the results determined can have on theparticular field taken into considaration and it does not allow retrieval of data based on the query entred by the user i.e. it has constrained scope.

In simple words, it works on static data rather than dynamic data. Unsupervised alghorithms like Vector Quantization are used for data compression, pattern recognition, facial and speech recognition, etc and therefore cannot be used in determining sentiment in twitter data.
Apriori algorithm fails to handle large datasets and as a result can generate faulty results.

Dis advantages:
The disadvantages of the project are as follows: ● The long retweets couldn't be retrieved fully, as a result it was represented by "...", so the
algorithm analyses it to be of neutral sentiment .
● Sarcasm was not detected in some sentences due to misuse of the semantics.
The proposed system architecture consists of five main steps. The first step for sentiment analysis of twitter data is to acquire the data that has to undergo sentiment analysis. Raw tweets were obtained at the end of this stage and stored in a database which was the input for preprocessing step.
Preprocessing consists of several steps which are the removal of URLs from the tweet, hashtag removal so that the tweet becomes more cleaned, removal of slangs, emoticon conversion to text, stop word removal.
Advantages:
The advantages of the project are as follows:
● The number of tweets were increased upto 1 Lakh to get better analysis of result whereas
tweepy could only retrieve 100 tweets per page.
● Analysing it using the populations and seat method gave an accuracy of nearly 90%.
● The website was deployed in heroku server for public use, which can be used for getting
sentiment analysis for any phrase or word.
● The twitter search api could retrieve data only of past 7 days.
● Giving hashtags for shorthand representation of the party gave ambiguous results in some
cases.
● Cannot get 100% accuracy in analysing the tweets.

V. PROP**OSED SYSTEM:**

The Proposed system is a browser which is completely related to online system, which provides the centralized database. It stores Defects data and description of the particular Defect data. It can also create reports and documents based on the information in its database.

Defect management is crucial to closing the loop between requirements, implementation and verification and validation. Traditional defect tracking management, implemented in a standalone method, can no longer address the complexity and pace of change in modern software development. Defect management processes must be strongly interlinked with all of the other software

ADVANTAGES:
● The number of tweets were increased upto 1 Lakh to get better analysis of result whereas tweepy could only retrieve 100 tweets per page.
● Analysing it using the populations and seat method gave an accuracy of nearly 90%.
● The website was deployed in heroku server for public use, which can be used for getting sentiment analysis for any phrase or word.
● The twitter search api could retrieve data only of past 7 days.
● Giving hashtags for shorthand representation of the party gave ambiguous results in some cases.
● Cannot get 100% accuracy in analysing the tweets.
● Cannot get 100% accuracy in analysing the tweets.

RESULTS FOR MODI AND RAHUL GANDHI TWEETS
Another way to analyze the whole prediction scenario is to compare the tweets between the leaders of the major national parties participating in the elections. This way the major sentiment of the masses towards the leaders that they are going to choose can be predicted easily.
This was performed by running the twitter sentiment analysis program with two search phrases: "Narendra Modi", for BJP, and "Rahul Gandhi", for Congress. The results for the two searches are as follows:

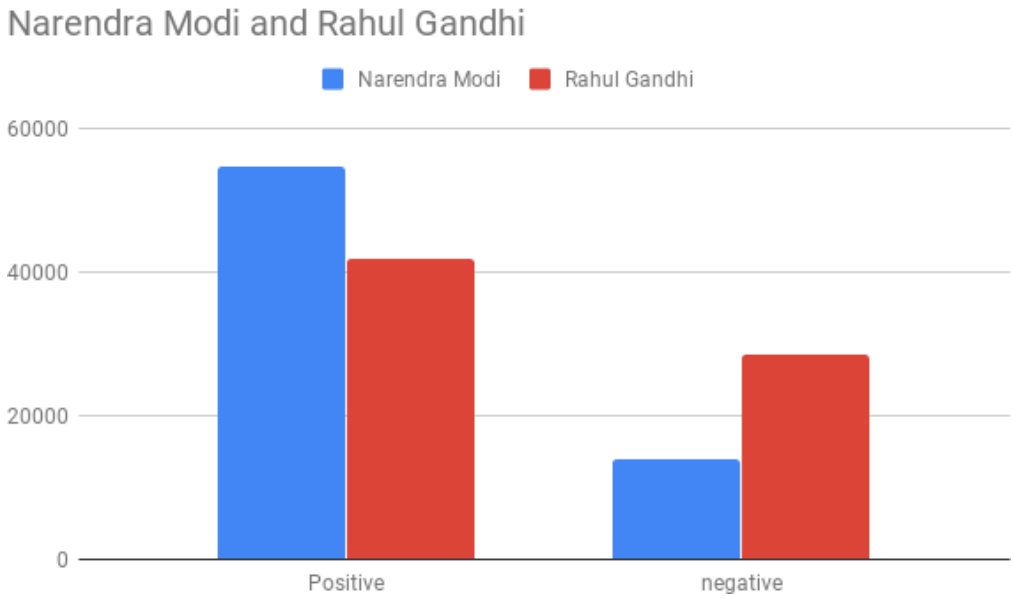


Figure 6: Comparison of sentiment analysis for Narendra Modi and Rahul Gandhi

As can be seen in the diagram, the competition is neck-to-neck between the two leaders of the major parties. However, Narendra Modi has slightly more positive and slightly less negative tweets, tipping the scale of favor towards Narendra Modi. This method however does not take into account all the different names used to reference the two.

VI. CONCLUSION

We took two ways to predict the results of elections

1. Using the population criteria

- As the variation by taking the ratios of positive to negative tweets was almost same. We tried considering the population factor into account for generalization purpose. We manipulated the total number of tweets as well as positive and negative tweets by dividing them with the current population percentage of the state.
- The results derived from it showed some deviation for the national parties and the regional parties of some states.
- The regional parties having less number of followers showed negligible tweets about them as a result the positive and negative tweets about them was unclear.

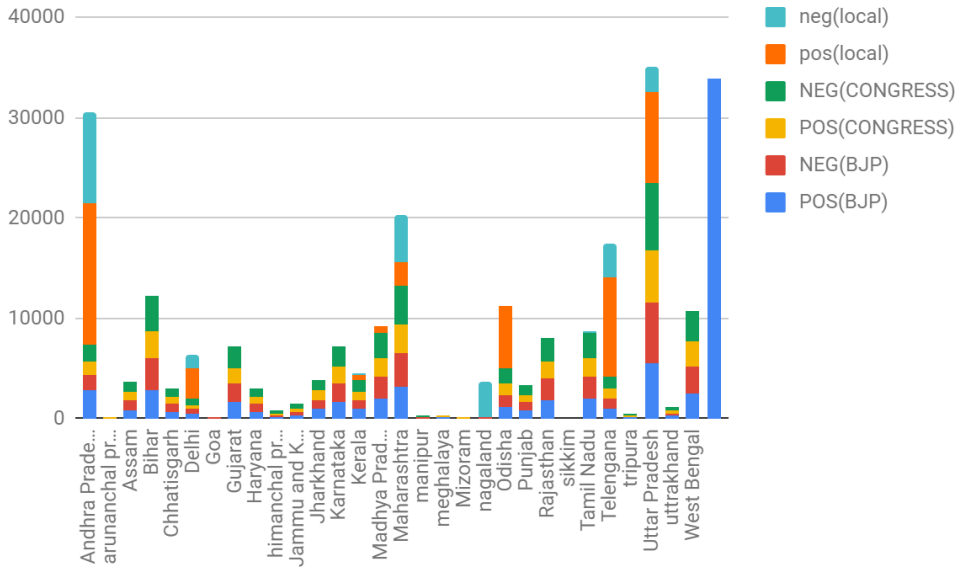


Figure 5: Positive and negative tweets comparison for regional parties as well as national parties

So we can see that,

1. States like Andhra Pradesh, West Bengal, Odisha, Tamil Nadu and Telangana have more positive sentiments as compared to negative sentiments so we can assume that winning party will be the local parties.
2. States like Gujarat, Goa, Madhya Pradesh, Uttar Pradesh and Uttarakhand have an inclination towards BJP government, giving them the winning points from those states.
3. States like Rajasthan, Haryana and Assam have a majority of congress.

According to the graph we can take a total number of positive tweets of bjp to be 33811 and congress to be 32110.As we are calculating the winning in the lok sabha election and the major parties contesting are BJP and congress we can clearly see that BJP is winning over the people's hearts and according to that BJP wins the lok sabha election.

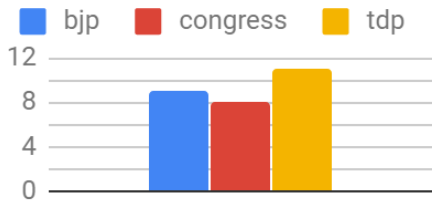
But there are certain drawbacks when we consider this method of calculation, which are discussed as follows:

1. When we consider the total population we are not considering the people eligible to vote.
2. There are many people whose view might change during election.
3. There are states with so less population that there vote affecting is negligible in this method.

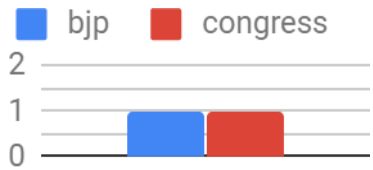
So, we chose to do the analysis in another process to get better and accurate results.

2. Using the seats in Lok Sabha

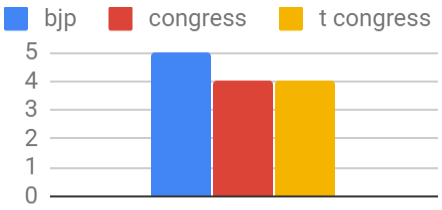
- The number of Lok Sabha seats are 543 which is contributed by each state. Representative from each state is elected for these seats. Party getting more than 271 get a majority to form the government or else parties form coalitions to getthe simple majority .
- So, the total tweets was taken and the positive tweets was assumed to be as seats for the party (calculating the percentage of positive tweet and multiplying it with the seats a single states contribute).Then we calculated it for individual states and the local parties.
- Even if the local parties win in the states they affect the total seats in the Lok Sabha according to the seat percentage which gave a much clearer idea about which party would win how many seats in the election.
- The results of each state is shown below :



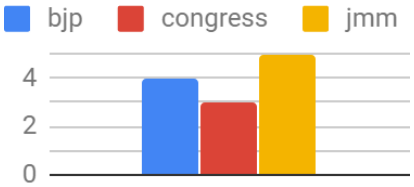
Andra Pradesh(TDP wins)



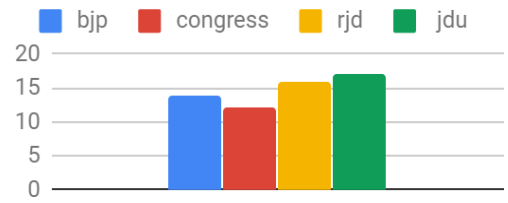
Aruachal Pradesh



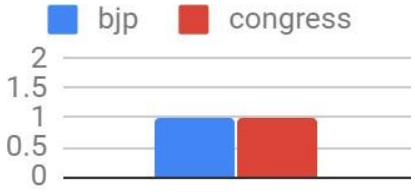
Assam(BJP wins)



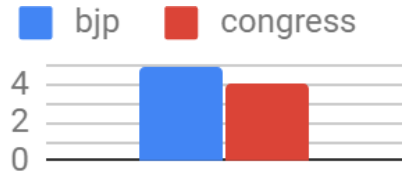
Chhattisgarh(JMM wins)



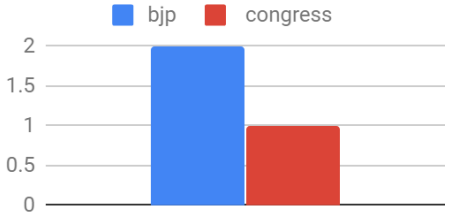
Bihar (JDU wins)



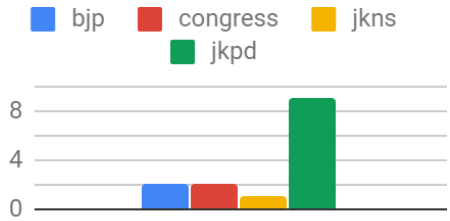
Goa (BJP=1, Congress=1)



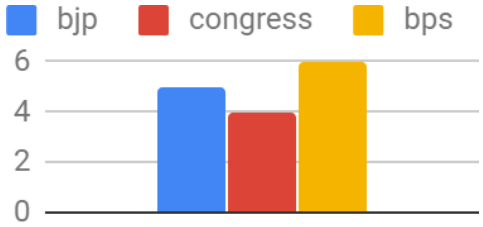
Haryana(BJP wins)



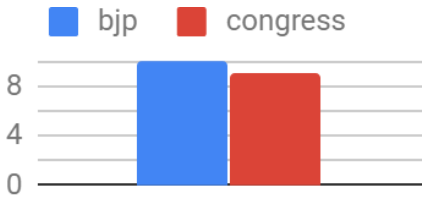
Himachal Pradesh(BJP wins)



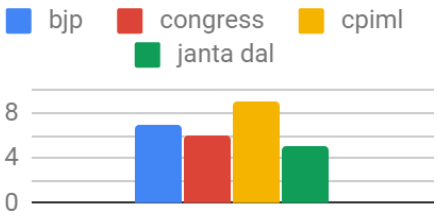
Jammu and Kashmir(JKPD wins)



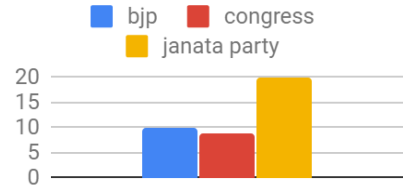
Jharkhand(BPS wins)



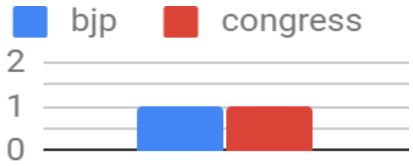
Karnataka(BJP wins)



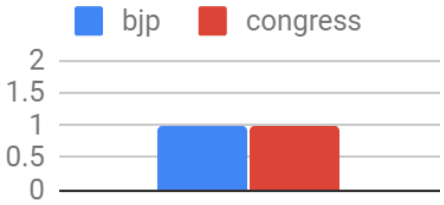
Kerela(CPIML wins)



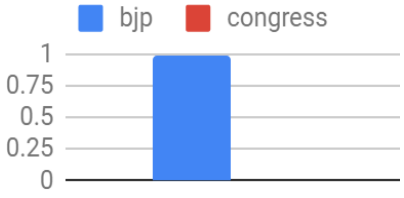
Madhya Pradesh(janta party wins



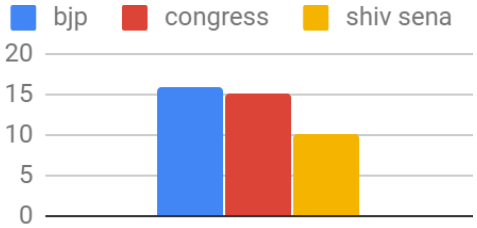
Meghalaya



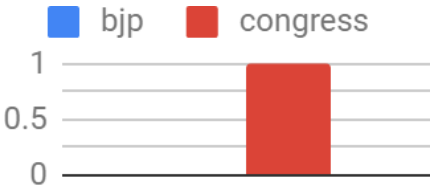
Manipur



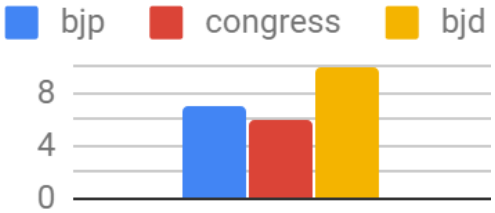
Mizoram(BJP wins)



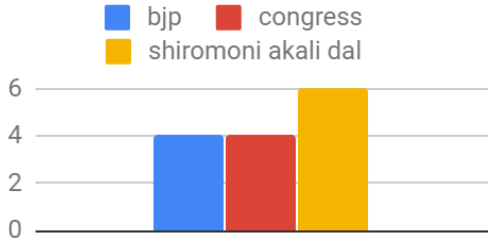
Maharashtra(BJP wins)



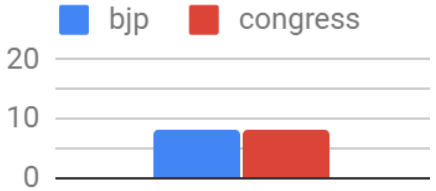
Nagaland(Congress wins)



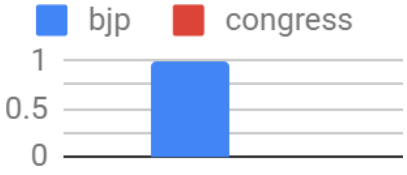
Odisha(BJD wins)



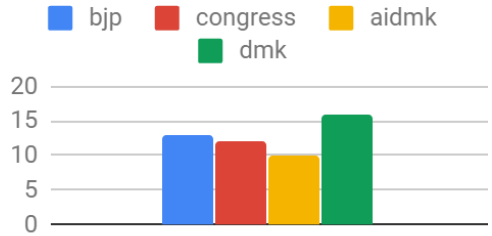
Punjab(SAD wins)



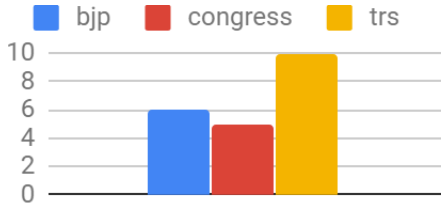
Rajasthan



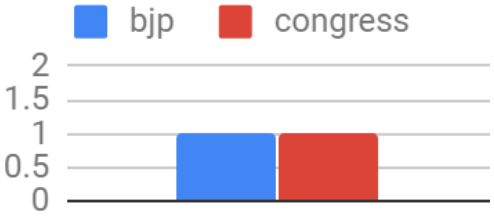
Sikkim(BJP wins)



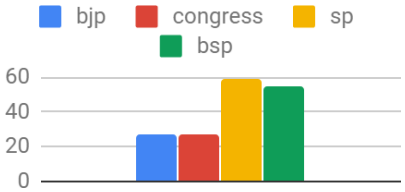
Tamil Nadu(AIDMK wins)



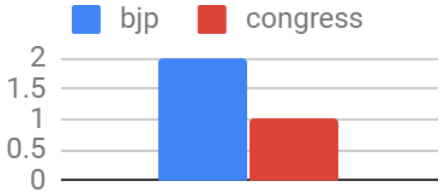
Telengana(TRS wins)



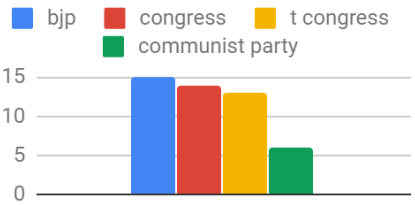
Tripura



Uttar Pradesh(SP wins)



Uttrakhand(BJP wins)



West Bengal(BJP wins

- From the value obtained by normalizing these data as can be seen from the graph we find out that *BJP wins 178 seats* in Lok Sabha and *congress wins 168 seats* in Lok Sabha. The parties have alliances with other local parties which is not taken into consideration in our project.
- If we see from the above data obtained and calculate we get that *BJP has more number of seats* so it would win the Lok Sabha elections but to have a simple majority it needs to align which other parties.

REFERENCES

[1]. Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls, By - Parnian Kassraie1, Alireza Modirshanechi and Hamid K. Aghajan, Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran, Islamic Republic of imec, Department of Telecommunications and Information Processing, University of Gent, Gent, Belgium.

[2]. Twitter Based Election Prediction and Analysis, By Pritee Salunkhe, Sachin Deshmukh, Department of Computer Science and Information Technology, Dr.Babasaheb Ambedkar Marathwada University, Maharashtra, India.

[3]. Election result prediction using Twitter sentiment analysis(IEEE), By Jyoti Ramteke, Samarth Shah, Darshan Godhia, Aadil Shaikh

[4]. Natural language processing(IEEE), By A. Gelbukh

[5]. Loper, Edward & Bird, Steven. (2002). NLTK: the Natural Language Toolkit. CoRR. cs.CL/0205028. 10.3115/1118108.1118117.