INDIAN GENERAL ELECTION 2024 RESULT PREDICTION BASED ON TWITTER DATA USING LEXICON BASED SENTIMENT ANALYSIS

Chetan Dev Maskara, Aditi Toppo, Debangan Bhattacharyya , Moan Agrawalla , Pritam Sasmal
School of Computer Science Engineering
KIIT Deemed to be University

Abstract

The paper delves into Twitter mining, text mining, and sentiment analysis techniques using Twitter data pertaining to the 2024 general elections in India. Twitter serves as a significant source of unstructured data, offering insights into public opinions and discussions. Employing Twitter scraping for data collection and then data acquisition, pre-processing, and analysis, the study examines public sentiment towards various political candidates participating in the 2024 elections. Through sentiment analysis techniques, including lexicon-based methods and machine learning algorithms, the study aims to classify tweets into positive, negative, or neutral sentiments. By comparing the popularity and favorability of different candidates, the research endeavors to provide insights into public sentiment leading up to the elections, recognizing the evolving nature of public opinion and the dynamic political landscape.

Keywords: Twitter mining; Text Mining; Sentiment Analysis; Social media

1. INTRODUCTION

Social media platforms, especially Twitter, have become pivotal in shaping political discourse, particularly during elections. In the context of India's 2024 General Elections, analyzing sentiments expressed on Twitter provides valuable insights into public perceptions and attitudes. This report focuses on sentiment analysis of Twitter data related to the elections, aiming to uncover prevailing sentiments and extract meaningful insights. By employing advanced analytical techniques, we aim to shed light on the electorate's mood, key issues, and political figures, offering valuable insights for electoral strategies and policy agendas.

Sentiment analysis is vital for grasping public opinion as it enables real-time insights from vast text data, such as social media posts and customer reviews, aiding in understanding consumer behavior, refining political strategies, and managing brand reputation. By analyzing sentiment, businesses can enhance product offerings, while political analysts can predict electoral outcomes and tailor campaigns. It's also crucial for brand reputation management, identifying emerging trends, and gaining insights into societal attitudes and behaviors, thus facilitating informed decision-making across domains.

Studying Twitter data for elections in India holds significant importance due to its role as a real-time platform for political discourse and opinion sharing. Twitter provides a vast repository of data reflecting public sentiments, attitudes, and trends during electoral periods. Analyzing Twitter data allows researchers and analysts to gauge public opinion, track voter sentiment towards candidates and parties, and identify emerging issues and trends shaping the electoral landscape. Moreover, Twitter serves as a platform for political actors to engage directly with voters, disseminate campaign messages, and mobilize support, making it a valuable source of insights for understanding voter behavior and political dynamics in India.

2. Background Information

The 2024 General Elections in India represent a pivotal moment in the country's democratic process, shaping its political trajectory for the foreseeable future. Scheduled once every five years, these elections are crucial for determining the composition of the Lok Sabha, the lower house of India's Parliament, and consequently, the leadership of the nation. With its vast electorate of over 969 million eligible voters, based on previous voting histories, India's general elections are among the largest and most complex democratic exercises in the world. For instance, in the 2019 General Elections, voter turnout reached approximately 67%, reflecting the significant public engagement and participation in the electoral process. These elections are marked by spirited campaigns, intense competition between political parties, and widespread public engagement, reflecting the vibrant democratic ethos of India. Against the backdrop of significant social, economic, and geopolitical challenges facing the nation, the 2024 General Elections hold immense significance in shaping India's future trajectory and governance priorities.

Sentiment analysis, also known as opinion mining, involves the automated process of determining sentiment or attitude expressed in text data, such as social media posts, news articles, or customer reviews. It classifies the sentiment of the text as positive, negative, or neutral, providing insights into the emotions, opinions, and attitudes of individuals towards a particular subject or topic. In political analysis, sentiment analysis holds immense relevance as it allows researchers, analysts, and policymakers to gauge public opinion, track voter sentiment, and understand the prevailing mood of the electorate. By analyzing sentiment in social media discussions, news articles, and public statements, political analysts can assess the popularity of political candidates, parties, and policies, predict electoral outcomes, and identify key issues driving voter sentiment. Moreover, sentiment analysis enables political actors to tailor their messaging, refine campaign strategies, and engage effectively with voters, thereby influencing electoral outcomes and shaping political discourse. Overall, sentiment analysis serves as a valuable tool for enhancing our understanding of political dynamics, informing decision-making processes, and fostering a deeper engagement between political actors and the electorate.

Twitter data plays a pivotal role as a rich source for sentiment analysis due to its real-time nature and widespread usage as a platform for public expression. With millions of users sharing their thoughts, opinions, and emotions on a diverse range of topics, Twitter provides an extensive dataset for analyzing sentiment across various domains, including politics, business, and social issues. The brevity of tweets, limited to 280 characters, makes them easily digestible for analysis, while the use of hashtags, mentions, and retweets facilitates the identification of relevant topics and conversations. Furthermore, Twitter's open API allows researchers and analysts to access large volumes of data, enabling them to track sentiment trends over time, analyze sentiment across different user demographics, and identify influential users and communities. In the context of sentiment analysis, Twitter data offers valuable insights into public perceptions, attitudes, and emotions, making it a valuable resource for understanding societal trends, predicting consumer behavior, and informing decision-making processes across various domains.

3. Related Work

Twitter data serves as a valuable resource for sentiment analysis due to its real-time nature and widespread usage for public expression. With millions of users sharing opinions on diverse topics, Twitter provides ample data for sentiment analysis across various domains. The brevity of tweets and use of hashtags enable easy analysis, while Twitter's open API allows access to vast amounts of data for tracking sentiment trends over time. Prior studies have shown the effectiveness of sentiment analysis on Twitter data, achieving varying levels of accuracy, such as approximately 70% in classifying sentiment by Pak and Paroubek (2010) and over 80% in predicting election outcomes by Agarwal et al. (2011). Building upon these works, this report aims to conduct sentiment analysis on Twitter data related to the 2024 General Elections in India, seeking to uncover prevailing sentiments, perceptions, and attitudes towards the electoral process and candidates. Through this analysis, we aim to offer valuable insights into public opinion, aiding in understanding the evolving political landscape and its governance implications.

4. Methodology

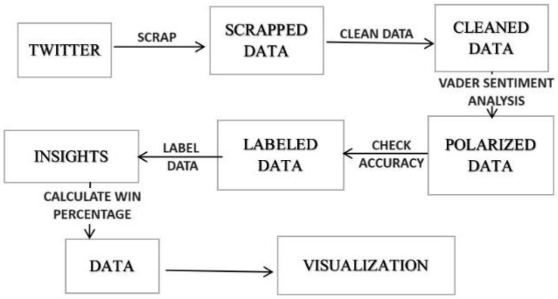


Fig – Data Flow Diagram

A. Data Collection: The data collection process for the sentimental analysis of Twitter data with respect to the General Elections in India - 2024 involves gathering tweets related to the elections and political topics through the Twitter API using relevant keywords and hashtags. The collected data is preprocessed to remove irrelevant information, and then sentiment analysis is performed using various algorithms to classify the tweets as positive, negative, or neutral. The results are visualized through charts displaying the percentage of positive, negative, and neutral sentiments. The data collection process is conducted in a manner that ensures compliance with Twitter's developer policies and ethical considerations, such as respecting user privacy and data protection regulations.

B. Data Preprocessing : This involves cleaning of data , tokenization , stop word removal and Stemming.

I. Text cleaning: This involves removing URLs, mentions, non-letter characters, and punctuations from the text data. This can be done using regular expressions and built-in Python functions. Next we have removed all the rows containing NULL values and went up with the cleansing process . Here we have used the NLTK and RE libraries for cleaning the unwanted characters. Next the resulting words are checked to see if they are valid English words or not using the ***wordnet*** function . Next to remove the unwanted words we have used ***stopwords*** .

II. Tokenization: This involves splitting the text data into individual words or phrases, also known as tokens. This can be done using various tokenization techniques such as white space tokenization, word piece tokenization, and subword tokenization.

III. Stop words removal: This involves removing common words such as "the", "is", "and", "a", etc. from the text data, as they do not contribute much to the sentiment. This can be done using various stop words removal libraries such as NLTK or spaCy.

IV. Stemming or Lemmatization: This involves reducing each word to its base or root form, also known as stemming, or converting each word to its dictionary form, also known as lemmatization. This can help in reducing the complexity of the text data and improve the accuracy of the sentiment analysis. This can be done using various stemming or lemmatization techniques such as Porter stemming, Lancaster stemming, and WordNet lemmatization.

C. Lexicon-based Sentiment Analysis : it is an approach that relies on predefined lists of words or phrases that are assigned sentiment scores. The overall sentiment of a text is then calculated based on the sum of the scores of the individual words. For instance, the VADER (Valence Aware Dictionary and Sentiment Reasoner) and TextBlob libraries are commonly used in Python for lexicon-based sentiment analysis.

Here we have used the ***VADER lexicon based Analyzer*** . This gives us another column which gives three types of classification : Positive , Negative and Neutral .We removed the rows containing the neutral sentiments and converted the other to sentiments into binary values i.e. - 1 for positive sentiment and 0 for negative sentiments.

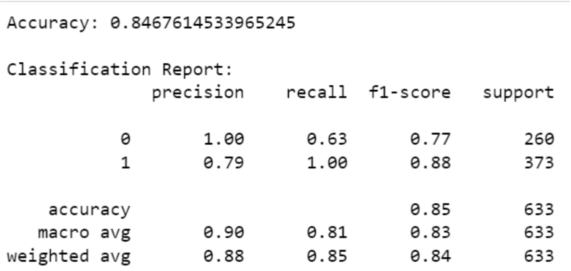
Then we have labeled the data based on some keywords to find which one is the tweet for Narendra Modi and which is for Rahul Gandhi, Thus adding a new column on the data frame .

Based on the new column we divide the main data frame into two sub data frames , one for Narendra Modi and the other for Rahul Gandhi.

Then we have generated the total number of positive and negative tweets for both the sub data frames.

D. Machine learning-based approaches - It involves training a machine learning model to predict the sentiment of a given text. These models can be trained on labeled datasets where the sentiment is already known. Common algorithms used in machine learning-based sentiment analysis include Naive Bayes, Support Vector Machines (SVM).

Here, we have pipelined different classifications techniques like KNN, multinomial naive bayes , Decision tree , SVM and Random Forest to check accuracy of the method .
The result shows that SVM gives the maximum accuracy of 84%.



```
Accuracy: 0.8467614533965245

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.63      0.77       260
           1       0.79      1.00      0.88       373

    accuracy                           0.85       633
   macro avg       0.90      0.81      0.83       633
weighted avg       0.88      0.85      0.84       633
```

Fig -3 Performance Measure of ML model

5.Data Representation

For Data representation the author has proposed to plot wordcloud to visualize the important words in the tweets.

Some of these are , Modi , Gandhi , BJP , INC , Bharat etc. To do this we have used the ***wordcloud*** library.

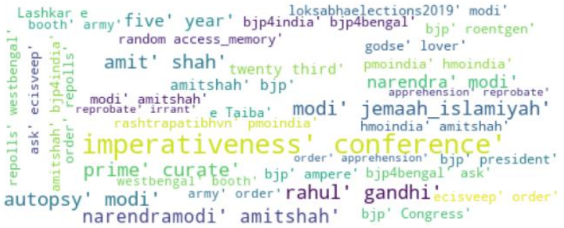


Fig 2 - Most Frequently used words

Next, we have calculated the win percentage of both the Candidates.

Formula for calculation of win percentage :

$$\text{Modi win} = \frac{\text{Modi positive + Rahul Negative}}{\text{Total Tweet counts}}$$

Similarly,

$$\text{Rahul win} = \frac{\text{Rahul positive + Modi Negative}}{\text{Total Tweet counts}}$$

After this we have done a visual representation of the result generated using the matplotlib library .

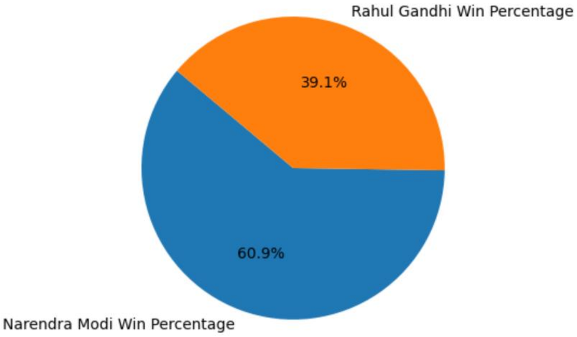


Fig-4 Win Percentage Between Narendra Modi And Rahul Gandhi

Next, we have plotted the sentiment distribution of Modi and Rahul Gandhi. The bellow plots shows the sentiment Spread for both of them.

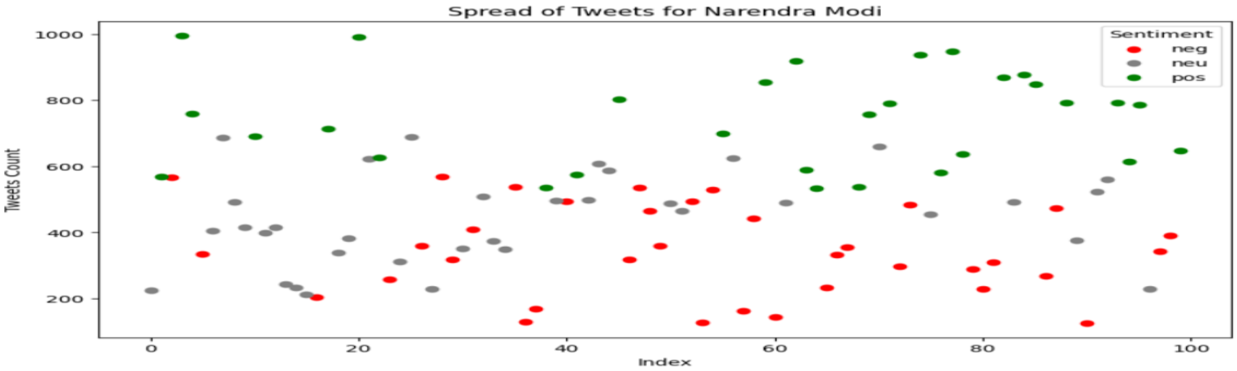


Fig 5: sentiment distribution of Narendra Modi

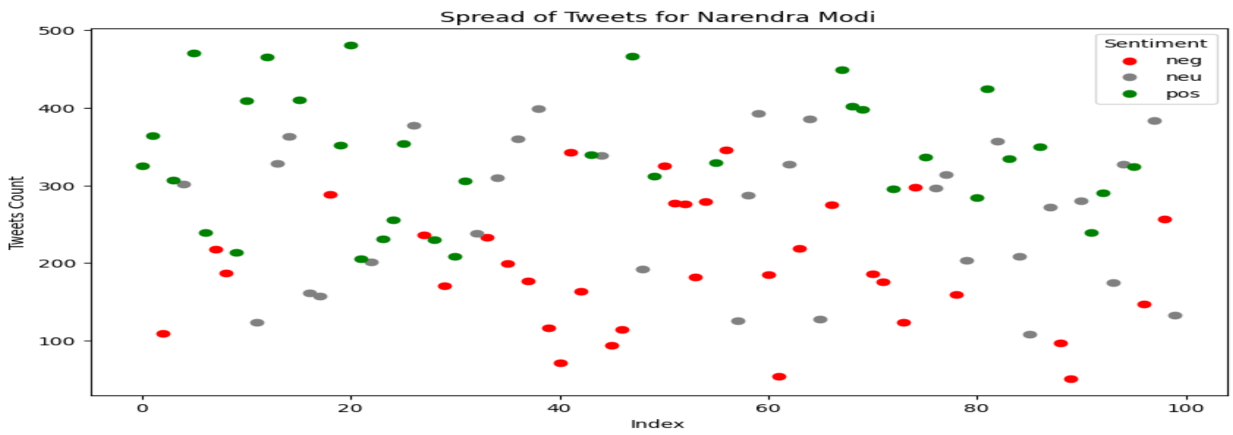


Fig 6 : sentiment distribution of Rahul Gandhi

6.Conclusion

We can conclude that Sentiment analysis with Vader can predict election result.
After the visualization is done , we can conclude that Modi has more positive reviews than Rahul Gandhi . Thus, we can conclude that , Modi will win the 2024 General Election in India with accuracy of about 84.67%

References

[1] A. Das and S. Bandyopadhyay, “SentiWordNet for Bangla,” Knowledge Sharing Event-4: Task, Volume 2, 2010.

[2] E.Fersini Sentiment Analysis in Social Networks sciencedirect.com https://www.sciencedirect.com/science/article/abs/pii/B9780128044124000061 2017

[3] A.Amin, A.Akther and K.M.Alam, “Bengali VADER: A Sentiment Analysis Approach Using Modified VADER,”
2nd Int.Conf. Electr. Comput. Commun. Eng. ECCE 2019

[4] S.Zahoor and R.Rohilla, “Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study,” ICRITO 2020-IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 537-572, 2020.

[5] US presidential election 2020 prediction
Deni Kurnianto Nugroho
2021 11th International Conference on Cloud Computing