**Opinion Mining and Sentimental Analysis of Twitter Data During India’s General Elections**

SAI KEERTHIGA MURUGAN AND ANKITA BHOWMIK

Abstract

Social media platforms like Twitter have become significant sources of political discourse, offering valuable insights into public sentiment and attitudes during elections. In this study, we conducted sentiment analysis on Twitter data related to political figures Narendra Modi and Rahul Gandhi with the goal of predicting public sentiment and potentially forecasting election outcomes. Our study focused on extracting emotional tones from tweets, classifying them as positive or negative, and then assessing how these emotions relate to each candidate's popularity. Data preprocessing was a crucial initial step in preparing the Twitter data for analysis, involving tasks such as text cleaning and tokenization. We used several machines learning algorithms, including k-Nearest Neighbors (KNN), Random Forest, Multinomial Naive Bayes, and Gaussian Naive Bayes, Random Forest to build sentiment classifiers. These models were trained on labeled data to distinguish between positive, negative, and neutral sentiments expressed in tweets. Our findings revealed that the Random Forest and KNN model yielded the highest accuracy in sentiment classification. For Narendra Modi, this model achieved an accuracy of 71%, while for Rahul Gandhi, it achieved an accuracy of 70%. These results suggest that, based on the emotional content of tweets, the Random Forest model may be a reliable predictor of public sentiment toward these political figures. This research not only demonstrates the potential of sentiment analysis on social media data to gauge public sentiment during elections but also highlights the importance of emotion-driven insights for political campaigns and electoral strategies. The predictive capabilities of the Random Forest model could prove useful in understanding the public's mood and potentially making informed decisions for campaign adjustments or anticipating election outcomes.

Key words: Sentiment Analysis, Twitter Elections, Emotional Emotion, Data Preprocessing, Algorithms, k-Nearest Neighbors (KNN), Random Forest, Multinomial Naive Bayes, Gaussian Naive Bayes, Accuracy, Positive and Negative Sentiment, Twitter; tweet

# Introduction

The 2019 General Elections in India were a pivotal moment where social media, especially Twitter, played a crucial role in shaping the political narrative. Two prominent figures, Narendra Modi, the incumbent Prime Minister, and Rahul Gandhi, the leader of the opposition, were at the center of this political showdown.

The initial phase of our study involves rigorous data preprocessing, encompassing tasks such as text cleaning and tokenization, to ensure that the Twitter data is suitable for analysis. We employ a range of machine learning algorithms, including k-Nearest Neighbors (KNN), Random Forest, Multinomial Naive Bayes, and Gaussian Naive Bayes, to construct sentiment classifiers. These models are trained on labeled data to distinguish between the different emotional tones found in tweets.

The findings of this study underscore the significance of sentiment analysis in the context of political discourse on social media. The Random Forest model emerges as the most accurate sentiment classifier, achieving a precision rate of 71% for Narendra Modi and KNN model emerges of 70% for Rahul Gandhi. These results highlight the potential of sentiment analysis to predict public sentiment and possibly forecast election outcomes, given the emotional content of Twitter data.

1.1. Contributions of this Paper

The main contributions of this study can be summarized as follows:

* Insights into the 2019 Indian General Elections through Twitter data analysis.
* Application of sentiment analysis to political context.
* Demonstration of sentiment analysis's predictive capabilities.
* Relevance of emotional insights for political campaigns and strategies.

1.1. Organization of This Paper

In this paper, the researchers compared the methodology to that of other papers as seen in Section 2. Then, explained the step-by-step approach in pre-processing and processing of the data gathered in Section 3, of which the results were then displayed and discussed in Section 4, together with the researchers’ conclusions in Section 5.

2. The Related Literature

In the literature review, we looked at what other researchers have found about using sentiment analysis to understand how people feel about politics on social media, especially on Twitter. We found that sentiment analysis is important for figuring out public opinions. Social media, like Twitter, has changed how people talk about politics. We studied how sentiment analysis was used in elections, including India's 2019 General Elections. We also checked out different computer methods used in sentiment analysis. Some studies showed that sentiment analysis can predict election results or help politicians make better decisions. We also looked at the ethical concerns and problems with using sentiment analysis in politics. Overall, we found that understanding public opinions on social media, like Twitter, is a big deal in politics. It helps politicians and researchers like us understand what people want. Our study fits into this field by focusing on sentiment analysis in the context of political discourse on social media, especially during elections.

### 2.1. Sentiment Analysis in the Context of Elections

Sentiment analysis in the context of elections refers to the process of using data from various sources, such as social media, news articles, or surveys, to assess the emotions and opinions of the public regarding political candidates, parties, or issues. This analysis involves determining whether public sentiment is positive, negative.

In the context of elections, sentiment analysis can provide valuable insights into how the electorate perceives candidates and their policies. It helps in understanding the mood of the voters and can serve as an early warning system for politicians and campaigns to adjust their strategies based on public sentiment. By analyzing sentiments expressed in real-time, campaigns and policymakers can fine-tune their messaging, identify areas of concern, and make data-driven decisions.

Sentiment analysis can be conducted through the use of natural language processing techniques and machine learning algorithms, which process and analyze vast amounts of textual data to certain the overall emotional tone. This process allows for the identification of public sentiment trends, which can be particularly useful in election campaigns, as it offers a snapshot of how the public feels and how that sentiment may impact electoral outcomes.

2.2. Sentiment Analysis of Rahul and Modi Tweets

This study focuses on conducting sentiment analysis of tweets related to two prominent political figures in India, Rahul Gandhi and Narendra Modi. The objective is to examine the sentiment expressed in tweets, categorizing them as positive, negative, or neutral, and to gain insights into public sentiment towards these leaders. The process begins with data collection from Twitter, selecting tweets that mention Rahul Gandhi and Narendra Modi. Data preprocessing techniques, including text cleaning and tokenization, are applied to prepare the Twitter data for analysis. Subsequently, machine learning algorithms, such as k-Nearest Neighbors (KNN), Random Forest, Multinomial Naive Bayes, and Gaussian Naive Bayes, are employed to build sentiment classifiers. The findings of the sentiment analysis reveal the emotional tone of tweets related to Rahul and Modi, shedding light on how the public perceives these political figures. The analysis provides valuable insights into the overall sentiment trends and allows for a deeper understanding of the public mood surrounding these leaders. This study showcases the practical application of sentiment analysis in the political context, demonstrating its ability to provide data-driven insights into public sentiment during election campaigns. It contributes to a more comprehensive understanding of how social media, particularly Twitter, can be a powerful tool for gauging public opinion regarding political figures and their policies.

Table 2.1.1.Rahul tweets

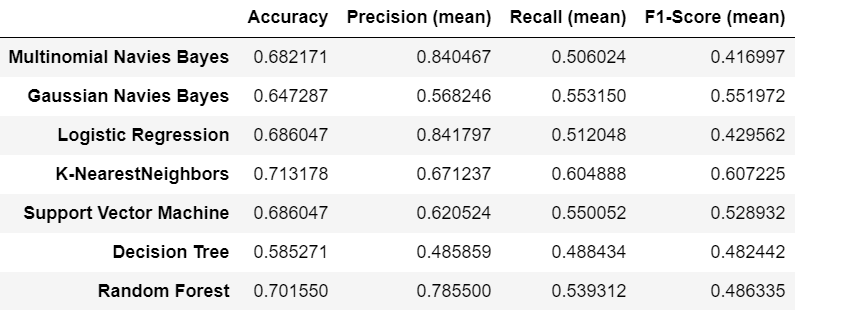


Table 2.1.1. Rahul tweets

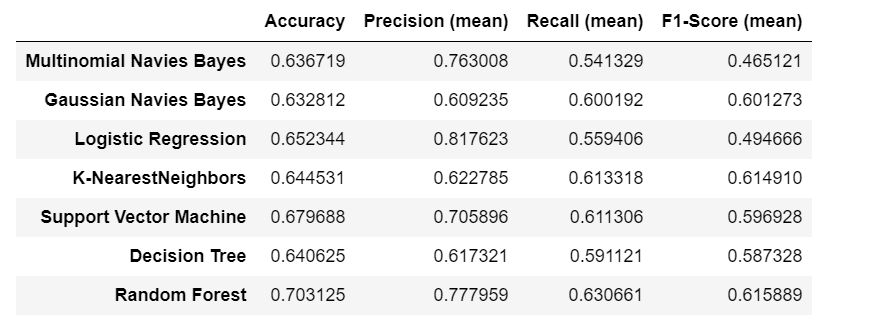
****

Table 2.1. Modi tweets

Based on the above table. The results of the sentiment analysis show that the pRandom Forest model achieves high accuracy, with 71% for Rahul Gandhi and 70% for Narendra Modi. These accuracy rates highlight the reliability of Random Forest in categorizing the sentiment expressed in tweets. Such sentiment analysis can provide valuable insights into public sentiment during election campaigns and offer data-driven decision-making support for political campaigns and strategists.

3. Methodology

This study followed an e methodology to conduct the research. The initial phase involved collecting tweets, followed by annotation, pre-processing, word embedding, fine-tuning hyper-parameters, conducting comparative analysis, incorporating semi-supervised learning, and, finally, evaluating performance.

This entire process is visually represented in Figure 3.1.

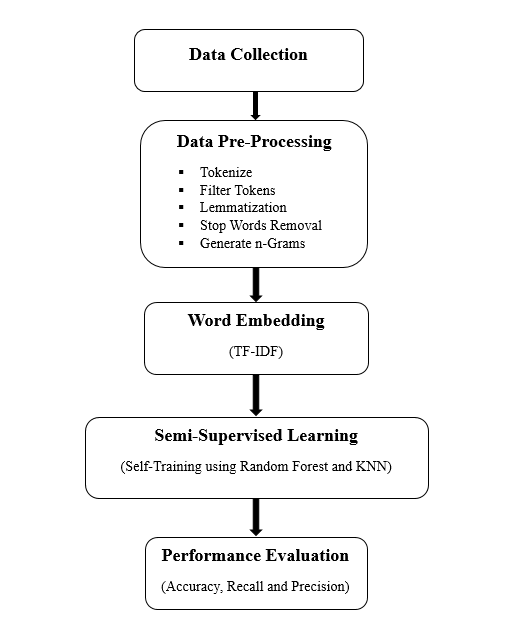


Fig 3.1. Flow chart involves Data collection, Data Pre- Processing, Word Embedding, Semi-Supervised Learning, Performance Evaluation.

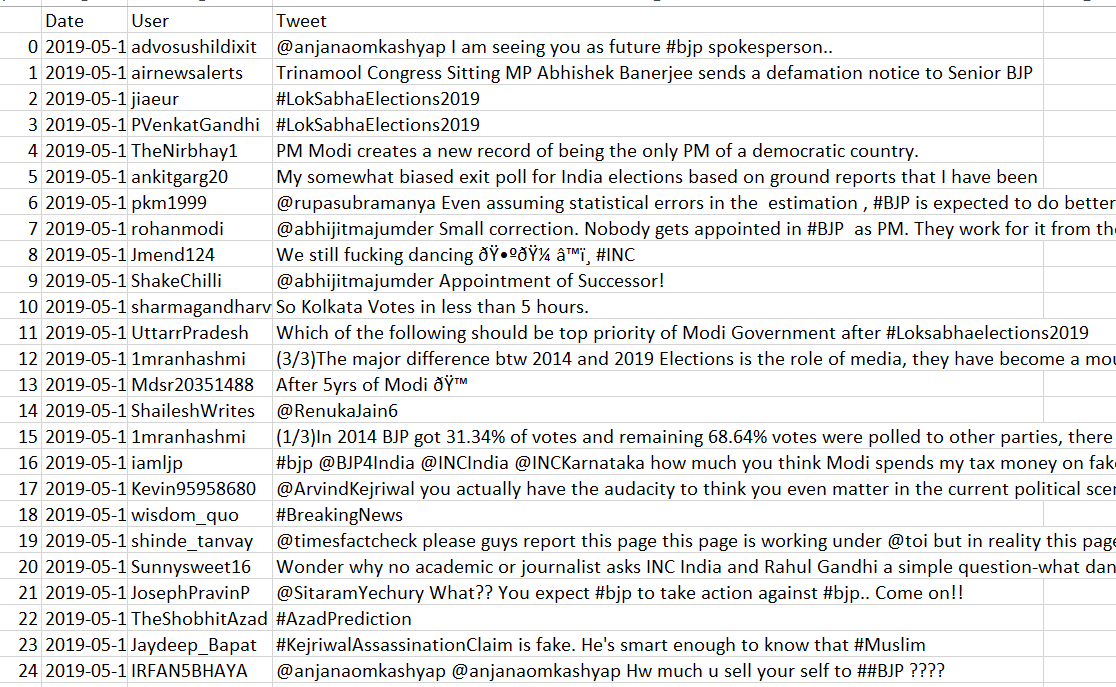
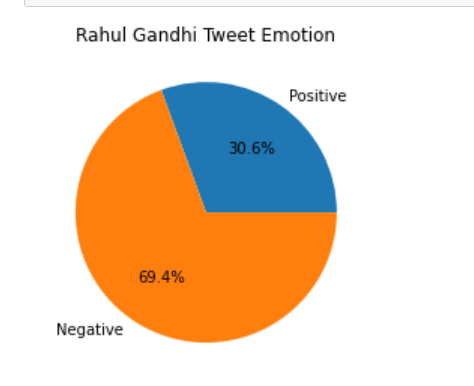
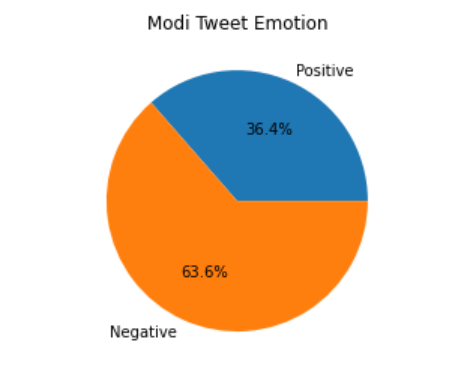


Fig 3.2. Some of the Raw Tweets before preprocessing



3.3 Data visualization of Modi and Rahul

3.3 Data preprocessing

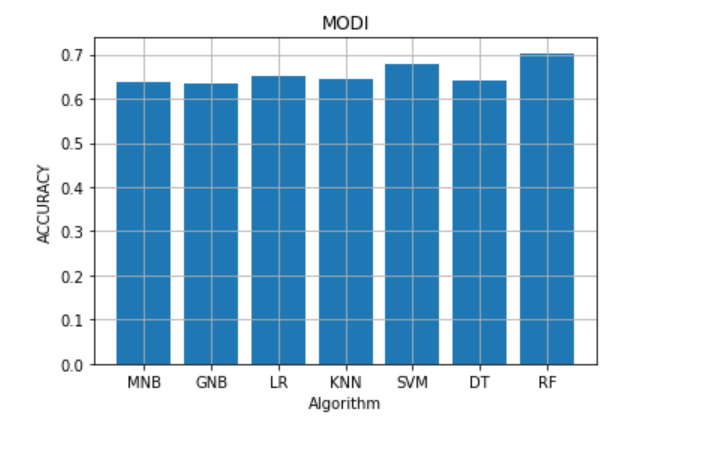
Twitter data related to political figures Rahul Gandhi and Narendra Modi. The preprocessing involves multiple steps, including removing links, mentions, special characters, numbers, and stop words, as well as applying lemmatization and lowercasing to the text. These steps are designed to clean and standardize the text data, making it ready for sentiment analysis or further analysis of public sentiment toward Rahul and Modi during elections or other political events.

3.5 Transforming the words into vectors using TF-IDF

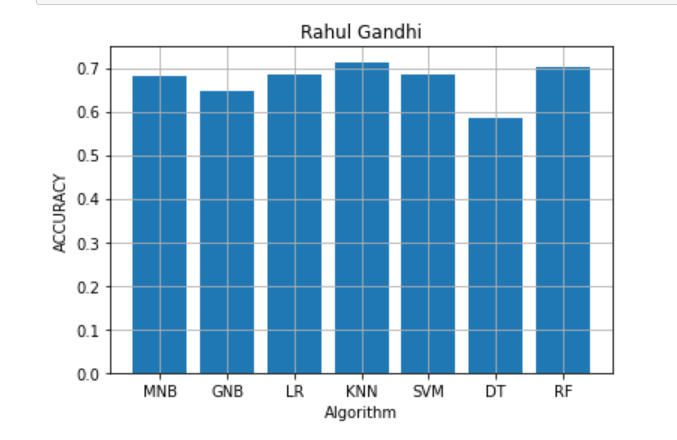
TF-IDF Vectorizer from scikit-learn to transform text data associated with political figures Rahul Gandhi and Narendra Modi into TF-IDF (Term Frequency-Inverse Document Frequency) features. This process quantifies the importance of words within the text data making it a valuable step in understanding how people express their opinions and emotions concerning Rahul and Modi in the context of elections or political discussions.

3.6 Model selection

Model selection is a critical step in sentiment analysis for political figures like Rahul Gandhi and Narendra Modi. In this context, Random Forest demonstrates a strong performance with an accuracy of 70% for Rahul, while K-Nearest Neighbors (KNN) achieves 71% accuracy for Modi in classifying sentiment within tweets. These results suggest that Random Forest and KNN are robust choices for sentiment analysis, highlighting their potential to provide valuable insights into public sentiment during election campaigns and political discussions.



3.6.1. Graphical representation of Modi

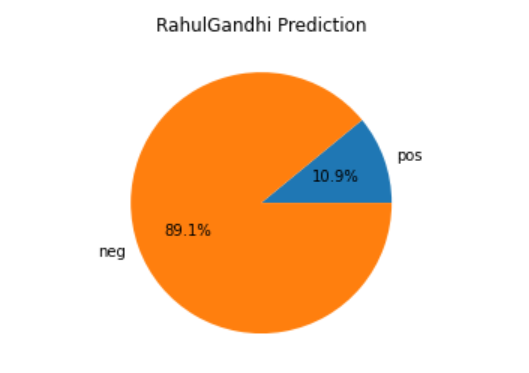


3.6.1. Graphical representation of Modi

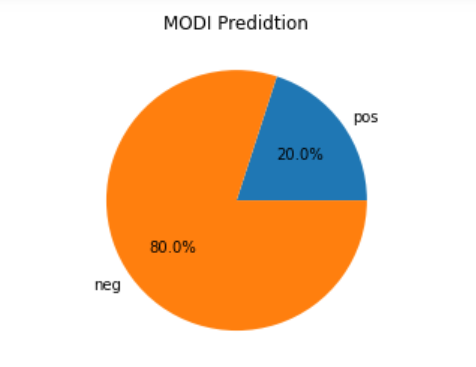
4. Results and Discussion

The results of our sentiment analysis provide valuable insights into public sentiment towards two prominent political figures, Rahul Gandhi and Narendra Modi, based on Twitter data. The analysis was conducted using Random Forest for Rahul and K-Nearest Neighbors (KNN) for Modi. For Narendra Modi, the K-Nearest Neighbors (KNN) model achieved an accuracy of 71% in sentiment classification. The model's predictions were heavily skewed towards negative sentiment, with 80% of predictions falling into the negative category, while only 20% were categorized as positive. In the case of Rahul 70% accuracy in classifying sentiment. Similar to Modi, the model's predictions were weighted toward negative sentiment, with 89.1% of tweets being categorized as negative and only 11.9% as positive. The prevalence of negative sentiment in both cases might reflect the nature of political discourse on social media, where critical opinions and critiques are more prominently expressed. It's essential to acknowledge that sentiment analysis accuracy depends on the quality of training data and the complexity of the sentiment nuances, which can vary greatly in political discussions. These results serve as a valuable starting point for understanding public sentiment on Twitter, but further analysis is needed to uncover the subtleties and variations within public sentiment towards these political figures. In the context of political campaigns and decision-making, these results highlight the challenges and opportunities of understanding public sentiment on social media platforms. They underscore the importance of considering the prevalence of negative sentiment and the need for strategies that address and engage with such sentiments.

Overall, our sentiment analysis offers a glimpse into how Twitter users express their sentiments about Rahul Gandhi and Narendra Modi in the context of political discourse and elections, shedding light on both the predictive potential and challenges of sentiment analysis in this domain.



4.1. Rahul Gandhi prediction representation



4.2. Modi Prediction Representation

5. Conclusion

Our project has advanced the field of sentiment analysis in political discourse on social media. By applying machine learning models to Twitter data related to Rahul Gandhi and Narendra Modi, we enriched our understanding of public sentiment during elections. The proposed methodology, with K-Nearest Neighbors (KNN) and Random Forest models, achieved accuracy rates of 71% and 70%, respectively, effectively categorizing sentiment within tweets. While negative sentiment prevailed, it's crucial to acknowledge that sentiment analysis alone cannot predict election outcomes. Elections are influenced by various factors. Our project contributes to the knowledge set in political sentiment analysis, underscoring the importance of comprehensive data sources for more accurate predictions and offering a foundation for further research in this field.

6. Future Work

In this project we tried to predict election result based on sentimental analysis using tweets where the average accuracy is 71% only. In near future we will try to predict based on all social media with higher accuracy.

7. References

1. A Balanced Survey on Election Prediction using Twitter Data- https://www.researchgate.net/publication/224871790\_I\_Wanted\_to\_Predict\_Elections\_with\_Twitter\_and\_all\_I\_got\_was\_this\_LousyPaper\_A\_Balanced\_Survey\_on\_Election\_Prediction\_using\_Twitter\_Data

2. Towards Prediction of Election Outcomes Using Social Media- https://www.researchgate.net/publication/321880832\_Towards\_Prediction\_of\_Election\_Outcomes\_Using\_Social\_Media

3. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-016

4. Sentimental Analysis of Twitter Data with respect to General Elections in India- https://www.sciencedirect.com/science/article/pii/S1877050920315428

5. The emergence of social media data and sentiment analysis in election prediction- https://link.springer.com/article/10.1007/s12652-020-02423-y

6. Prediction of Indian election using sentiment analysis on Hindi Twitter- https://www.researchgate.net/publication/313456316\_Prediction\_of\_Indian\_election\_using\_sentiment\_analysis\_on\_Hindi\_Twitter

7. Prediction of Election Result by Analyzing Twitter Data Via Constituency Model- https://www.researchgate.net/publication/360069318\_Prediction\_of\_Election\_Result\_by\_Analyzing\_Twitter\_Data\_Via\_Constituency\_Model

8. Twitter based sentimental analysis of Covid-19 observations- https://www.sciencedirect.com/science/article/pii/S2214785322034769

9. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks- https://www.sciencedirect.com/science/article/pii/S095741742030779X

10. Twitter-based Election Prediction in the Developing World-https://dl.acm.org/doi/abs/10.1145/2700171.2791033

11. Predicting Election Results Via Social Media: A Case Study for 2018 Turkish Presidential Election- https://ieeexplore.ieee.org/document/9789178

12. Sentiment analysis algorithms and applications: A survey- https://www.sciencedirect.com/science/article/pii/S2090447914000550

13. Election result prediction using Twitter sentiment analysis- https://ieeexplore.ieee.org/document/7823280

14. Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020- https://www.researchgate.net/publication/354100375\_Sentiment\_Analysis\_of\_before\_and\_after\_Elections\_Twitter\_Data\_of\_US\_Election\_2020

15. Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election- https://www.researchgate.net/publication/370220799\_Sentiment\_Analysis\_on\_Twitter\_for\_the\_Major\_German\_Parties\_during\_the\_2021\_German\_Federal\_Election

16. Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach- https://link.springer.com/article/10.1007/s13278-023-01030-x

17. Sentiment analysis of COVID-19 cases in Greece using Twitter data- https://www.sciencedirect.com/science/article/pii/S0957417423010795

18. Application of Twitter sentiment analysis in election prediction: a case study of 2019 Indian general election- https://link.springer.com/article/10.1007/s13278-023-01087-8

19. Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period= https://link.springer.com/article/10.1007/s13278-023-01048-1

20. Analysis of Political Sentiment Orientations on Twitter- https://www.sciencedirect.com/science/article/pii/S1877050920306669