Sentiment Analysis of Indian General Election Tweets

***Abstract*—The use of social media platforms has revolutionized the way people express their opinions and engage in political discussions. In this report, we present a comprehensive analysis of sentiment expressed in tweets during the Indian General Election. By leveraging sentiment analysis techniques, we aimed to gain insights into public sentiment towards political parties and candidates during this crucial democratic process. Our study involved collecting a large dataset of tweets, preprocessing the data, performing sentiment analysis, and analyzing the results. The findings shed light on the prevailing sentiment patterns and public opinion surrounding the election, providing valuable insights for political analysts, policymakers, and candidates.**

***Index Terms*—Sentiment Analysis, Linear SVC, TF-IDF, Political Tweets**

I. Introduction

As social media has become a popular platform for people to express their opinions and share information, it has also become a significant source of data for analyzing public sentiment and perception towards political leaders, parties, and policies. Sentiment analysis is a technique that involves using machine learning algorithms to classify text data into positive, negative, or neutral sentiments. The sentiment analysis of tweets related to the Indian General Elections can provide valuable insights into the mood of the people towards different political parties and leaders, the effectiveness of their campaigns, and the key issues that are important to the voters.

In this report, we present an analysis of sentiment in Indian General Election tweets using a dataset obtained from Kaggle [1]. The dataset covers the period from February to May 2019, offering a valuable glimpse into public sentiment during that time. The primary objective of this study is to train a sentiment analysis model and evaluate its performance in predicting sentiment based on the textual content of the tweets.

II. Dataset and Preprocessing

The dataset used for this analysis was sourced from Kaggle [1]. Initially, a different dataset, known as Sentiment140, was utilized to train the sentiment analysis models. However, to test the models' performance on Indian General Election tweets, the original dataset mentioned in the report's description was used.

The preprocessing steps involved cleaning the tweets to remove stopwords, emojis, URLs, and other irrelevant elements and performing tokenization, stemming, and lemmatization. This process is crucial for eliminating noise and enhancing the accuracy of sentiment prediction.

III. Models and Training

Three machine learning models were trained on the Sentiment140 dataset after the preprocessing steps to understand the frequency distribution of words, the most common words used, and the sentiment distribution of the tweets. The models used for sentiment analysis were:

1. Linear SVC (Support Vector Classifier)
2. Logistic Regression
3. Multinomial Naive Bayes

These models were chosen for their effectiveness in text classification tasks. Each model was trained using the cleaned tweets and their corresponding sentiment labels.

IV. Performance Evaluation

After training the models, their performance was evaluated using both training and testing datasets. The accuracy scores were calculated to assess their predictive capability. The Linear SVC model achieved the highest accuracy, with a training accuracy of 88% and a testing accuracy of 77%.

Application to Indian General Election Tweets:

To evaluate the sentiment of Indian General Election tweets, the trained Linear SVC model was applied to the original dataset. The tweets in this dataset were cleaned using the same preprocessing steps as the Sentiment140 dataset.

By running queries on the dataset, sentiment labels were obtained for tweets related to specific keywords such as “BJP,” “Modi,” “Congress,” etc. This allowed for identifying positive and negative sentiments associated with these keywords.

V. Libraries Used:

1. Pandas: Used for data manipulation and

analysis.

2. NumPy: Used for numerical operations and array manipulation.
3. Scikit-learn: Utilized for machine learning tasks, including model training, evaluation, and preprocessing.
4. NLTK (Natural Language Toolkit): Employed for text preprocessing tasks such as removing stop words and emojis.
5. Regular Expressions (regex): Used for pattern matching and text cleaning tasks.

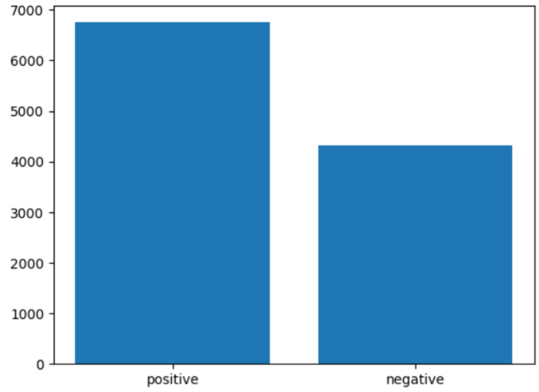
VI. Results

A query function was developed to analyze the sentiment distribution for specific keywords within the Indian General Election tweets dataset. For example, a query was run to analyze the sentiment associated with the keyword "BJP." The function extracted tweets containing the keyword applied the trained sentiment analysis model (Linear SVC) and categorized the sentiments as positive or negative.

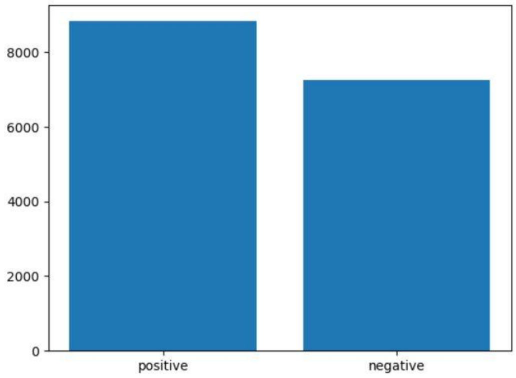
The results were visualized using a graph, providing an overview of the sentiment trends related to the BJP during the election period. This graph depicted the proportions of positive and negative sentiments expressed in tweets mentioning the BJP, offering insights into the public perception of the party.

The query function can be used to analyze sentiments for other keywords such as "Modi" or "Congress," allowing researchers and analysts to understand public sentiment towards different aspects of Indian politics.

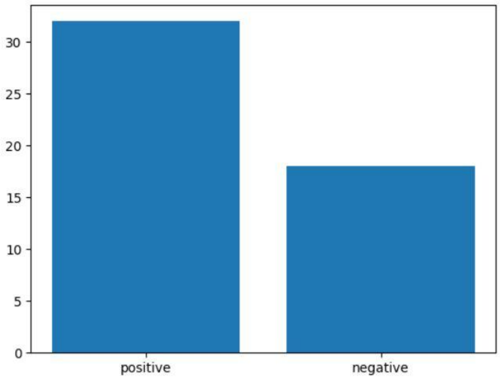
By tracking sentiment trends, stakeholders can identify shifts in public opinion, evaluate campaign effectiveness, and assess sentiment towards specific policies or political figures. These insights support data-driven decision-making and strategy development.



Graph displaying the number of positive & negative tweets for “Bjp”



Graph displaying the number of positive & negative tweets for “Congress”



Graph displaying the number of positive & negative tweets for “Sisodia”

It is important to note that the sentiment analysis results obtained through the query function are estimations based on the trained model's predictions.

In conclusion, the query function proved to be a valuable tool for analyzing sentiment distribution in Indian General Election tweets. The graphical representation provided a clear understanding of public sentiment trends, enabling stakeholders to make informed decisions and adapt strategies accordingly.

Future improvements could include analyzing sentiment in multi-word phrases or incorporating sentiment intensity analysis for nuanced sentiment expressions. Additionally, expanding the dataset to include more recent election cycles would enhance the understanding of public sentiment dynamics in Indian politics.

VII. Evaluation

The pipeline was trained using the training dataset and corresponding sentiment labels. After training, the model was used to predict the testing data. The predicted results were then compared with the actual results using evaluation metrics such as the confusion matrix and the classification report, which provides an accuracy score for the model.

The confusion matrix helps us understand the performance of the sentiment analysis model by showing the counts of true positive, true negative, false positive, and false negative predictions. The classification report provides a comprehensive evaluation of the model's performance, including metrics such as precision, recall, F1-score, and support for each sentiment class.

The accuracy score obtained from the classification report serves as a measure of the model's overall performance in predicting sentiment. A higher accuracy score indicates a better alignment between predicted and actual sentiment.

VIII. Conclusion

In this report, we conducted sentiment analysis on Indian General Election tweets using a dataset sourced from Kaggle. The sentiment analysis models were trained using the Sentiment140 dataset, and their performance was evaluated. The Linear SVC model demonstrated the highest accuracy in predicting sentiment.

Applying the trained model to the original Indian General Election tweets dataset provided insights into the sentiment associated with specific keywords. This analysis can help understand public sentiment towards political parties and candidates during election campaigns. It highlights the importance of sentiment analysis in shaping campaign strategies, identifying key issues, and addressing voter concerns effectively.

Moreover, our findings demonstrate the potential of social media analytics in understanding public opinion and sentiment in real time. By leveraging the power of machine learning algorithms and natural language processing techniques, we can analyze massive amounts of social media data in a matter of hours, providing us with valuable insights that would have been impossible to obtain through traditional survey methods.

Future work could involve exploring more advanced natural language processing techniques, experimenting with other machine learning algorithms, and incorporating additional features to improve sentiment prediction accuracy. Furthermore, extending the analysis to include more recent election cycles would provide valuable insights into evolving sentiment trends.

IX. References

[1] Kaggle Dataset: Indian Political Tweets 2019 (Feb to May) Sample. Retrieved from: https://www.kaggle.com/datasets/codesagar/indian-political-tweets-2019-feb-to-may-sample

[2] https://www.kaggle.com/datasets/kazanova/sentiment140

[3] Documentation:scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[4] Documentation:scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html