Prediction of Indian Election using Sentiment Analysis on Twitter(X) Data

Archana Suhas Vaidya, PhD
Department of Computer Engineering Assistant Professor
GES's R. H. Sapat College of Engineering Management Studies and Research, Nashik, India

Nida Shaikh
Department of Computer Engineering Research Scholar
GES's R. H. Sapat College of Engineering Management Studies and Research, Nashik, India

Dipak V. Patil, PhD
Department of Computer Engineering Professor
GES's R. H. Sapat College of Engineering Management Studies and Research, Nashik, India

ABSTRACT

Opinion mining, also known as sentiment analysis, involves classifying subjective sentiments in text into three categories: positive, negative, and neutral. This process delves into words and phrases to uncover the emotions expressed within sentences, paragraphs, or entire documents. Effective sentiment analysis relies on thorough text preprocessing, including the removal of irrelevant characters, stop words, and punctuation, as well as tokenization, which breaks text into meaningful units while preserving contextual relationships.

When applied to Twitter (X) data, sentiment analysis faces unique challenges due to the informal language, abbreviations, hashtags, and extraneous characters often present in tweets. Comprehensive preprocessing is essential to address these issues, enabling the extraction of meaningful insights. Among the various methods available, Long Short-Term Memory (LSTM) networks excel in sentiment analysis because of their ability to capture sequential and contextual nuances inherent in textual data.

This work proposes leveraging an LSTM-based model to perform opinion mining on Twitter (X) data related to the Indian election. The proposed model demonstrates exceptional performance, achieving an F1 score of 97.47% and an accuracy rate of 97.78 %. These results highlight not only the robustness of the LSTM approach but also its superiority in outperforming competing models for sentiment analysis tasks.

Keywords

Sentiment analysis, Opinion mining, LSTM, Twitter data, Indian elections, machine learning, public opinion, opinion mining

1. INTRODUCTION

The utilization of Long Short-Term Memory (LSTM) in predicting Indian elections through sentiment analysis on Twitter (X) data marks a groundbreaking shift in discerning public sentiment and prognosticating electoral results. As the prevalence of social media platforms like Twitter continues to surge, millions of users actively voice their opinions and viewpoints on political candidates, parties, and pertinent issues. This wealth of user-generated content presents an unparalleled opportunity to gauge the prevailing mood among the electorate. Leveraging sentiment analysis methodologies, facilitated by tools like Textblob, enables a systematic analysis of this data, unveiling the spectrum of positive, negative, and neutral sentiments associated with various political entities. Through the aggregation and interpretation of these sentiments, researchers can glean valuable insights into public perception and potential voting behavior.

The fusion of machine learning (ML) and natural language processing (NLP) methodologies serves as a cornerstone in advancing the predictive capabilities concerning Indian elections through the analysis of social media discourse. ML algorithms are instrumental in discerning intricate patterns and emerging trends within the data, thereby facilitating more precise forecasts of electoral outcomes grounded in sentiment analysis. Moreover, these algorithms can undergo rigorous training to discern subtle nuances and contextual intricacies embedded in language, further refining the accuracy of predictive models. Complementary to ML, NLP techniques facilitate the processing and comprehension of textual data extracted from tweets, encompassing tasks such as entity recognition, topic extraction, and sentiment categorization. The amalgamation of ML and NLP empowers researchers to construct resilient predictive frameworks, offering real-time insights into the party-political landscape while accommodating regional idiosyncrasies and the dynamic influence of current events on voter sentiment.

The integration of sentiment analysis with LSTM, ML, and NLP necessitates a nuanced approach to address pertinent challenges, including data biases, the proliferation of misinformation, and ethical considerations encompassing user privacy. Striking a balance between harnessing the potential of social media data and mitigating associated risks is imperative in upholding the integrity and reliability of predictive analyses concerning Indian elections. Nonetheless, the convergence of LSTM with ML and NLP heralds a promising trajectory in comprehending and forecasting electoral dynamics, underpinning a data-driven approach towards enhancing political prognostication and democratic discourse.

2. LITERATURE REVIEW

In their study [1], the authors explore the intricate temporal and statistical dynamics of sentiment analysis within political campaigns, employing a comprehensive twofold analysis approach. Their investigation entails the collection of an extensive corpus comprising thousands of Twitter messages pertaining to political parties and leaders in the period preceding and succeeding the 2019 Spanish presidential election. The objective is to elucidate the nuanced shifts in sentiment dynamics prevalent in political discourse. To quantify these dynamics, the paper employs an array of statistical indices including entropy, mutual information, and the Compounded Aggregated Positivity Index, providing insights into alterations within the density function of sentiment data. Furthermore, manifold learning techniques such as auto encoders and stochastic embedding isdeployed to extract features from nonlinear intrinsic patterns embedded within the data. The findings underscore the informative nature of sentiment dynamics, elucidating discernible fluctuations in sentiment behavior and polarity across distinct political entities. These fluctuations arefound to be contingent upon various factors including the ideological stance of political parties, their geographical influence, and their alignment with nationalist or globalist ideologies. The study offers a comprehensive examination of how sentiment analysis serves as a potent tool for understanding political discourse during electoral cycles, offering valuable perspectives on the influence of sentiment dynamics on electoral outcomes.

Another study [2] emphasizes the significant effect of social media in transforming public debate and opinion, particularly during electoral campaigns. Social media platforms serve as conduits for political campaigns and mobilization efforts, ranging from peaceful advocacy to contentious confrontations. Previous research underscores the efficacy of sentiment analysis in harnessing user behavior and sentiments expressed online to forecast electoral trends. Recent advancements in natural language processing (NLP) and machine learning (ML), notably with the advent of long short-term memory (LSTM) models and bidirectional encoder representations from transformers (BERT), have revolutionized the landscape of sentiment analysis. Notably, studies focusing on the US 2020 presidential elections have highlighted the efficacy of these sophisticated language models in Twitter sentiment analysis, with the BERT model accurately predicting Biden's victory. These findings underscore the viability of sentiment analysis as a potent tool for electoral prognostication, emphasizing the pivotal role of deep learning methodologies in unraveling the complexities of political landscapes.

The emergence of social media platforms as reservoirs of rich data has facilitated nuanced analyses of social phenomena and the anticipation of forthcoming events, including electoral outcomes. Leveraging the vast troves of user-generated content, researchers have endeavored to gauge public sentiment and forecast election results. However, this Endeavor is not without its challenges, as social media data inherently harbors biases and noise. Prior research [3] has identified limitations in simplistic methodologies such as raw tweet counting, which often yield skewed results due to inadequate representation of the electorate. To address these challenges, scholars have proposed refined techniques such as weighted counting based on geographical factors and sociological similarities between regions. Demonstrations in various electoral contexts, including Turkey's 2018 presidential election, have underscored the superiority of domain-specific information and weighted counting strategies over conventional polling methods. By mitigating data biases and leveraging location-based insights, researchers have made significant strides in harnessing social media data for robust election predictions. Future endeavors in this domain aim to augment prediction models by integrating advanced machine learning techniques and deeper NLP analyses, thereby enhancing the capacity to capture linguistic nuances and contextual sentiments across diverse demographic spectra.

3. PROPOSED SYSTEMS

The primary aim of this study is to construct a robust model capable of accurately discerning positive or negative sentiments from arbitrary inputs. To fulfill this objective, a Long Short-Term Memory (LSTM) model is proposed, leveraging its inherent capacity for precise sentiment classification. To ensure optimal performance, it is imperative to select an appropriate dataset. Figure 1 explains proposed system architecture. In this study, a dataset sourced from Twitter was chosen for sentiment analysis. After data acquisition, meticulous data preparation was conducted to ready it for processing through the neural network model. This entailed comprehensive data preprocessing procedures aimed at refining the dataset. Following this preparatory phase, the data was seamlessly integrated into the proposed LSTM model. Initially, the model underwent meticulous design and subsequent training on the Twitter dataset to enhance its analytical capabilities. Upon completion of the training phase, the model was subjected to rigorous testing using new input data to ascertain its proficiency in sentiment classification.

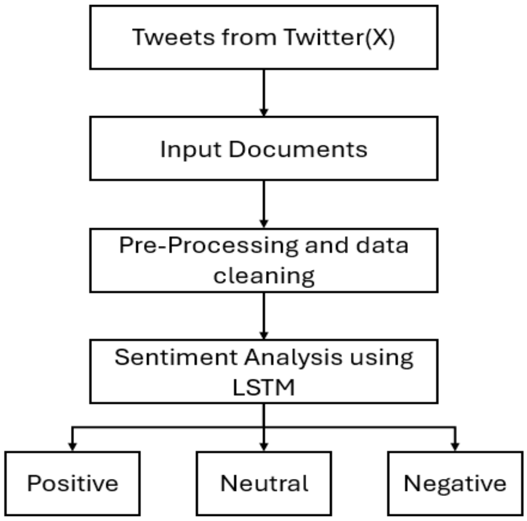

Figure 1 Proposed System Architecture for Twitter(X) Indian election tweets sentiment analysis using LSTM

The resultant output of this classification process delineates whether the input exhibits a negative or positive sentiment. Furthermore, the performance of the LSTM model was rigorously assessed using a battery of metrics to gauge its efficacy and reliability.

4. RESULT AND DISCUSSION

In addition to the model development process, an intuitive user interface (UI) is crucial for facilitating seamless interaction with the sentiment analysis tool. To accomplish this, Streamlit was selected, a user-friendly framework for UI development that allows for the creation of interactive web applications with minimal effort. Leveraging Streamlit's intuitive design capabilities, a user interface was crafted that enables users to input text data effortlessly and obtain sentiment analysis results in real-time. Furthermore, for insightful data visualization, the capabilities of both Matplotlib and Plotly were harnessed. These powerful libraries allowed for the generation of interactive plots and visualizations, providing users with a comprehensive understanding of sentiment trends and patterns within the analyzed data.

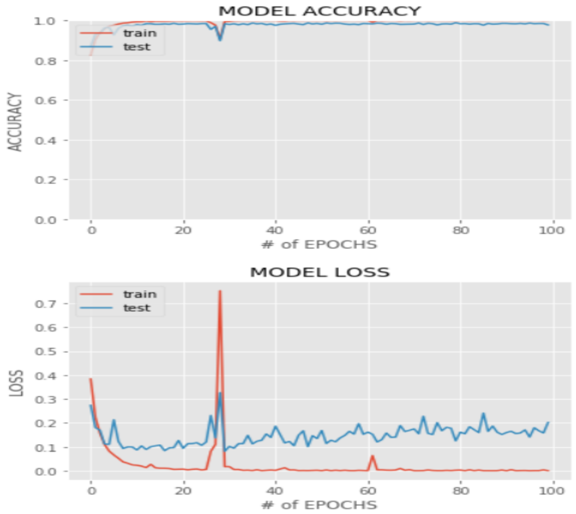

Figure 2 Model Accuracy and Loss for training data

Through the integration of Streamlit for UI development and Matplotlib along with Plotly for visualization, the user experience was enhanced, facilitating a more insightful exploration of sentiment analysis results.

User feedback mechanisms were implemented to continuously improve the tool based on user experience. By incorporating feedback loops, the UI evolved to meet the needs of its users. This iterative process not only refined the user interface but also expanded its functionality, making it more versatile and user-centric.

Figure 2 shows model accuracy and loss for training data. Preprocessing steps involve the elimination of stop words, symbols, and extraneous spacing to streamline the dataset. Subsequently, the dataset is partitioned into an 80:20 ratio for training and testing respectively, ensuring adequate representation for model evaluation. Through iterative training spanning 100 epochs, the model undergoes refinement to optimize its performance. Evaluation metrics reveal an impressive accuracy rate of 98% and a minimal loss of 0.20 for the test dataset, indicative of the model's efficacy in sentiment classification. Moreover, inferences drawn from the dataset suggest that the abundance of positive tweets correlates with electoral success, implying that political candidates or parties with a higher prevalence of positive sentiment are more likely to emerge victorious in elections.

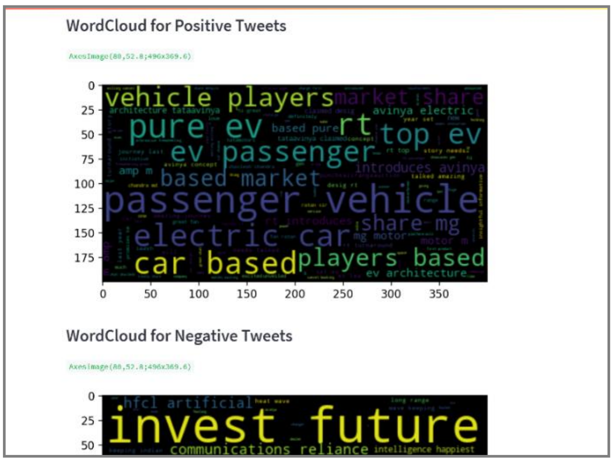

Figure 3 Word cloud using Seaborn library

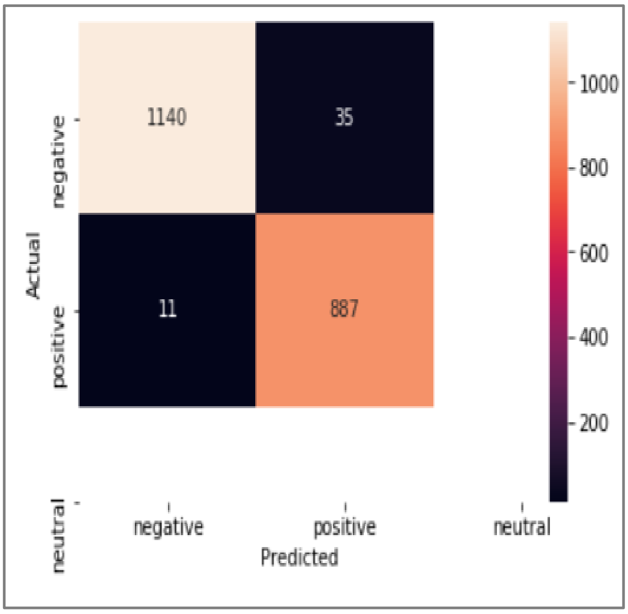

Figure 4 Performance evaluations with Confusion Metrics

- True Positives: 887(Predicted True and True in reality)
- True Negative: 1140(Predicted False and False in realtity)
- False Positive: 35 (Predicted Positve but Negative in reality)
- False Negative: 11 (Predicted Negative but Positive in reality)

Figure 3 shows word cloud using Seaborn library. When applied to Twitter (X) data, sentiment analysis encounters unique challenges stemming from the informal language, abbreviations, hashtags, and extraneous characters commonly found in tweets. To tackle these complexities and extract meaningful insights, comprehensive preprocessing is essential. Long Short-Term Memory (LSTM) networks have proven particularly effective in sentiment analysis due to their ability to capture sequential and contextual nuances in textual data.

Figure 4 shows LSTM-based model performance for opinion mining on Twitter (X) data related to the Indian election. The model's performance is evaluated using two key metrics: **Accuracy** and the **F1 Score**, defined as follows:

Accuracy measures the ratio of correctly classified samples to the total number of samples as per equation 1:

$$Accuracy = \frac{\text{Number of Correctly Predicted Samples}}{\text{Total number of Samples}} \quad \textbf{(1)}$$

The proposed model achieves an impressive Accuracy of 97.78%**.**

F1 Score represents the harmonic mean of Precision and Recall, balancing their trade-off as calculated in equation 2.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \textbf{(2)}$$

The proposed model demonstrates a remarkable F1 Score of97.47%, indicating a robust balance between identifying relevant sentiments and minimizing false positives or negatives.These metrics underscore the effectiveness of the LSTM model, establishing its superiority in addressing sentiment analysis tasks and outperforming competing approaches in this domain.

5. CONCLUSIONS AND FUTURE SCOPE

In conclusion, the endeavor encompassed the development of a robust sentiment analysis framework specifically tailored for political discourse, leveraging the dynamic capabilities of Long Short-Term Memory (LSTM) models. Through meticulous dataset selection, rigorous preprocessing, and model training, an LSTM-based system capable of discerning nuanced sentiment patterns within Twitter data pertaining to political campaigns was devised. This process involved selecting high-quality, representative datasets that accurately reflect the multifaceted nature of political discussions on social media. Rigorous preprocessing steps ensured that the data fed into the model was clean and meaningful, enhancing the accuracy and reliability of sentiment predictions.The integration of Streamlit for user interface development and Matplotlib alongside Plotly for visualization ensured a user-friendly and visually engaging platform for exploring sentiment analysis results. Streamlit's capabilities allowed for the creation of an interface that is not only easy to navigate but also interactive, providing real-time sentiment analysis results based on user input. Matplotlib and Plotly facilitated the creation of dynamic and informative visualizations, enabling users to grasp complex sentiment trends and patterns at a glance.By harnessing the power of deep learning and cutting-edge visualization techniques, this work contributes significantly to advancing the understanding of public sentiment dynamics during electoral cycles. The framework offers valuable insights into the intersection of social media discourse and political landscapes, shedding light on how public opinion evolves in response to political events and campaigns. This tool can serve as a critical resource for political analysts, campaign strategists, and researchers, enabling them to better understand and respond to the ever-changing sentiments of the electorate. Through this comprehensive approach, significant strides have been made in enhancing the methodological rigor of sentiment analysis in political contexts, while also making these insights accessible and actionable for a broader audience.

The future potential for expanding the capabilities of the sentiment analysis system developed in this research project is very promising. One avenue for future exploration involves integrating cutting-edge deep learning architectures, such as Transformer-based models like BERT or GPT, to further enhance the precision and depth of sentiment analysis outcomes. Additionally, the inclusion of multimodal data sources, such as images or audio, could enrich the analysis by capturing sentiments expressed through various mediums. Expanding the analysis to encompass multilingual sentiment analysis could broaden its applicability and provide insights into sentiments conveyed in diverse linguistic contexts. Furthermore, implementing real-time data streaming functionalities could facilitate continuous monitoring of sentiment trends, enabling timely insights into evolving public opinions. Finally, extending the application of the sentiment analysis framework beyond politics to diverse domains like marketing or customer feedback analysis could unlock novel applications and avenues for societal impact. By continually innovating and exploring these forward-looking directions, the sentiment analysis framework stands poised to evolve into a versatile and indispensable tool for comprehending and interpreting human sentiment across a spectrum of contexts.

6. REFERENCES

[1] M. Rodríguez-Ibáñez, F. -J. Gimeno-Blanes, P. M. Cuenca-Jiménez, C. Soguero-Ruiz and J. L. Rojo-Álvarez, "Sentiment Analysis of Political Tweets From the 2019 Spanish Elections," in IEEE Access, vol. 9, pp. 101847-101862, 2021, doi: 10.1109/ACCESS.2021.3097492.

[2] R. Chandra and R. Saini, "Biden vs Trump: Modeling US General Elections Using BERT Language Model," in IEEE Access, vol. 9, pp. 128494-128505, 2021, doi: 10.1109/ACCESS.2021.3111035.

[3] C. Bayrak and M. Kutlu, "Predicting Election Results Via Social Media: A Case Study for 2018 Turkish Presidential Election," in IEEE Transactions on Computational Social Systems, vol. 10, no. 5, pp. 2362-2373, Oct. 2023, doi: 10.1109/TCSS.2022.3178052.

[4] Balaji et al., 2016, M.S. Balaji, K.W. Khong, A.Y.L. Chong, Determinants of negative word-of-mouth communication using social networking sites Information & Management, 53 (4) (2016), pp. 528-540.

[5] S. N. Singh and T. Sarraf, "Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 112-116, doi: 10.1109/Confluence47617.2020.9058128.

[6] W. Zhao et al., "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 1, pp. 185-197, 1 Jan. 2018, doi: 10.1109/TKDE.2017.2756658.

[7] R. M. Duwairi and I.Qarqaz, "A framework for Arabic sentiment analysis using supervised classification" , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016.

[8] Shoukry, Amira & Rafea, Ahmed. (2012). Preprocessing Egyptian Dialect Tweets for Sentiment Mining.

[9] [Online Blog] Sentiment Analysis using NLTK – A Practical Approach, Sriniketh J — June 25, 2021, https://www.analyticsvidhya.com/blog/2021/06/sentiment-analysis-using-nltk-a-practical-approach/.

[10] L. Wasser and C. Farmer, "Sentiment Analysis of Colorado Flood Tweets in R", Earth Lab, 2018. [Online]. Available: https://earthdatascience.org/courses/earth-analytics/get-data-usingapis/sentiment-analysis-of-twitter-data-r/. [Accessed: 01- Mar- 2018].

[11] Sentiment Analysis using TextBlob, Parthvi Shah Jun 27, 2020, [Online]. Available: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

[12] How to build a real-time live dashboard with Streamlit, By AbdulMajedRaja RS, Posted in Community, April 21 2022[Online] Available at: https://blog.streamlit.io/how-to-build-a-real-time-live-dashboard-with-streamlit/.

7. AUTHOR'S PROFILE

Archana S. Vaidya received a Ph.D. degree in computer engineering from Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. She received her master's degree in computer engineering from V.J.T.I., Mumbai University (INDIA) in 2010 and bachelor's degree in computer engineering from Walchand College of Engineering Sangli Shivaji University (INDIA) in 2002. Her areas of interest are parallel computing and machine learning. She has teaching experience of 22 years. She is a life member of ISTE and IAENG. She has published 9 Scopus-indexed papers and 40+ journal papers.

Nida Shaikhreceived a B.E. degree in computer engineering in 2022 from SPPU, Pune, Maharashtra India and an M.E. degree in computer engineering in 2024 from SPPU, Pune Maharashtra, India.

Dipak V. Patil received a B.E. degree in computer engineering in 1998 from the University of North Maharashtra India and an M.Tech. degree in computer engineering in 2004 from Dr.Babasaheb Ambedkar Technological University, Lonere, India. He has done a Ph.D. degree from S.R.T.M. University, Nanded. Currently, he is a professor and head of the Computer Engineering Department at GES R.H. Sapat College of Engineering, Management Studies and Research, Nashik, India which is affiliated with the Savitribai Phule Pune University, Pune, and Maharashtra, India 422002.