ANALYZING POLITICAL SENTIMENT OF INDIC LANGUAGES WITH TRANSFORMERS

Pranav Gunhal

Artificial Intelligence Coalition, Sunnyvale, California, USA

ABSTRACT

This paper presents an analysis of sentiment on Twitter towards the Karnataka elections held in 2023, utilizing transformer-based models specifically designed for sentiment analysis in Indic languages. Through an innovative data collection approach involving a combination of novel methods of data augmentation, online data preceding the election was analyzed. The study focuses on sentiment classification, effectively distinguishing between positive, negative, and neutral posts, while specifically targeting the sentiment regarding the loss of the Bharatiya Janata Party (BJP) or the win of the Indian National Congress (INC). Leveraging high-performing transformer architectures, specifically IndicBERT, coupled with specifically fine-tuned hyperparameters, the AI models employed in this study achieved remarkable accuracy in predicting the the INC's victory in the election. The findings shed new light on the potential of cutting-edge transformer-based models in capturing and analyzing sentiment dynamics within the Indian political landscape. The implications of this research are far-reaching, providing invaluable insights to political parties for informed decision-making and strategic planning in preparation for the forthcoming 2024 Lok Sabha elections in the nation.

KEYWORDS

Sentiment analysis, Twitter, Karnataka elections, Bharatiya Janata Party, Indian National Congress, transformers, Indic languages, data augmentation, IndicBERT, political decision-making.

1. INTRODUCTION

Sentiment analysis has emerged as a critical field of research within Artificial Intelligence (AI), finding diverse applications in politics, social media, and market research [1]. The ability to capture and analyze public sentiment provides invaluable insights into individuals' opinions, attitudes, and emotions, empowering organizations and policymakers to make informed decisions. Among the plethora of social platforms, Twitter stands out as a prominent source of real-time, user-generated data reflecting public sentiment [2].

The Indian political landscape is a dynamic and vibrant arena where local sentiments play a pivotal role in shaping electoral outcomes [3]. In this context, sentiment analysis of political discourse on Twitter has proven to be a powerful tool for understanding public opinion, predicting election results, and devising effective political strategies, especially in Western elections [4].

Karnataka, located in the southern region of the Indian subcontinent, is a state known for its linguistic diversity, with over 150 languages spoken throughout its territory. The quinquennial Karnataka state elections held in May 2023 provide an ideal case study to explore the sentiment dynamics and predictive capabilities of Natural Language Processing (NLP) classification.
This paper aims to present a comprehensive analysis of sentiment on Twitter regarding the Karnataka elections, offering valuable insights into the sentiment dynamics leading to the victory of the Indian National Congress (INC) and the defeat of the current ruling party, Bharatiya

Dhinaharan Nagamalai (Eds): EMVL, EDUT, SECURA, AIIoT, CSSE -2023
pp. 143-156, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120512

Janata Party (BJP). Prior to the elections, the state was long under the control of the local Janata Dal (Secular) (JDS) party, followed by becoming a regional stronghold of theBJP in South India. Given the BJP's defeats in multiple other local elections throughout India, especially in the South, failure to win Karnataka in the 2023 state elections would be a significant setback for the ruling party. Such a loss could further impact the party's efforts in the upcoming national elections, Lok Sabha 2024, where the BJP seeks to maintain its status as the ruling national party.

Leveraging advancements in transformer-based models, specifically designed for sentiment analysis in Indic languages [5], this study contributes to the growing body of literature on sentiment analysis and political forecasting, focusing on the Indian political system. To achieve our research objectives, we adopted a multi-faceted approach that involved innovative data collection techniques, advanced transformer architectures, and targeted model optimization. Our methodology included web scraping to collect a diverse range of tweets, ensuring a representative sample of sentiments expressed in the period leading up to the election. To augment the dataset, we employed multiple data augmentation techniques to enhance its diversity and generalizability.

Building upon the success of transformer-based models in various natural language processing tasks [6], we utilized state-of-the-art architectures specifically tailored for sentiment analysis in Indic languages. The selected models, such as BERT [7] and IndicBERT [8], are known for their ability to capture semantic nuances, contextual information, and linguistic patterns inherent in Indian languages, making them highly suitable for analyzing sentiment in the context of Karnataka elections.

The significance of this research lies in its potential to provide political parties with timely and accurate insights into public sentiment, facilitating informed decision-making and strategic planning for the forthcoming 2024 Lok Sabha elections. By leveraging the power of NLP and sentiment analysis, political actors can gauge the effectiveness of their campaigns, identify potential voter concerns, and tailor their messaging to align with the prevailing sentiment [9].

This paper is structured as follows: Section 2 provides a comprehensive review of related work in sentiment analysis, political forecasting, and the use of Twitter data in analyzing public sentiment. Section 3 details the dataset collection process, including web scraping and data augmentation techniques employed. Section 4 presents the methodology, encompassing the transformer-based models, hyperparameter tuning, and evaluation metrics. Section 5presents the experimental results and analyses, highlighting the accuracy achieved in predicting the sentiment towards the BJP's defeat or the INC's victory. Finally, Section 6 concludes the paper and discusses avenues for future research.

2. LITERATURE REVIEW

2.1. Machine Learning for Sentiment Classification

Machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, and Maximum Entropy, have been widely employed for sentiment classification tasks [14]. Researchers have explored various features, including n-grams, syntactic patterns, and lexical resources, to improve the performance of sentiment analysis models [16]. Additionally, feature selection and dimensionality reduction techniques, such as Information Gain and Principal Component Analysis, have been utilized to enhance the efficiency and accuracy of sentiment classification [14].

Deep learning models, particularly neural networks, have demonstrated remarkable performance in sentiment analysis tasks. Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have been successful in capturing contextual information and sequential dependencies in textual data [14]. Convolutional Neural Networks (CNNs) have shown effectiveness in extracting local features and patterns from text [15]. Recently, transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [6], have gained prominence due to their ability to capture contextual information and semantic nuances in large-scale language modeling tasks, including sentiment analysis [6].

2.2. Political Forecasting and Twitter Data

The analysis of public sentiment on social media platforms, especially Twitter, has emerged as a powerful tool for political forecasting and understanding public opinion. Twitter data, with its vast volume and real-time nature, provides a valuable resource for capturing and analyzing public sentiment [10]. Various studies have explored the relationship between Twitter sentiment and real-world events, including elections, stock markets, and social phenomena.

The predictive power of Twitter sentiment in elections has been investigated by researchers. Tumasjan et al. [4] demonstrated that Twitter data can predict election outcomes by analyzing political sentiment expressed in tweets. Bollen et al. [2] revealed a correlation between Twitter mood and stock market fluctuations. Gohil and Vasanwala [3] conducted sentiment analysis on Twitter data to analyze public opinion on Indian politics.

Furthermore, the sentiment analysis of Twitter data has been employed to uncover public sentiment on specific political events, policies, and leaders. Researchers have utilized sentiment analysis to study public reactions to government initiatives, election campaigns, and political speeches [11]. The analysis of sentiment dynamics on Twitter provides insights into the shifting public opinion and sentiment fluctuations over time [12].

In summary, sentiment analysis has advanced significantly with the adoption of machine learning, deep learning, and transformer-based models. Twitter data has emerged as a valuable resource for understanding public sentiment, predicting election outcomes, and studying political dynamics. The combination of sentiment analysis techniques and Twitter data provides researchers and policymakers with valuable insights into public opinion and sentiment.

3. DATA COLLECTION AND AUGMENTATION

This section outlines the data set collection process and describes the data augmentation techniques employed for sentiment analysis. The data set was obtained through web scraping of Twitter data, and two augmentation techniques were applied: Random Swap and Language Transformation using Google Translate.

3.1. Data Collection and Analysis

The dataset for sentiment analysis was collected by employing web scraping techniques to extract tweets from the Twitter platform. The Twitter API was utilized to retrieve tweets based on specific keywords, hashtags, or user profiles, through the Tweepy library. The collected tweets were processed to ensure the removal of personally identifiable information, thus maintaining privacy and adhering to ethical guidelines. The full text of the tweet and the date-time stamp of the tweet were collected for analysis in this study. While the majority of the tweets collected werein English, a portion of them contained words and phrases in Kannada, as well as the Indian national language Hindi. These tweets were not changed or translated, as the study aims toclassify sentiment of multilingual tweets.

3.2. Random Swap

The Random Swap [18] technique involves randomly selecting two words within a sentence and swapping their positions. By repeating this process for a given number of iterations, the sentence structure is modified, generating new training examples. The Random Swap technique introduces variations in sentence structure, enabling the model to learn and generalize better by considering different word arrangements. This allowed us to augment the text to increase the robustness of thetraining data.

3.3. Language Transformation with Google Translate

To further diversify the data set, Language Transformation using Google Translate [19] was employed. This technique involves translating the original sentences into different languages,such as Kannada or Hindi, using the Google Translate API. The translated versions of the sentences were included in the augmented data set, enhancing its diversity and enabling themodel to handle multilingual inputs effectively. Integrating the translated sentences into the dataset enriches its linguistic diversity and enhances the models' ability to handle and interpretmultiple languages effectively.

Table 1. Examples of data annotations for each category

Label	**Tweet**
Yes	“They are there on Indian Passport not Karnataka Passport why mention kannadigas. You guys would nothave even given a heed if there were no elections in Karnataka. Shame on you CONgressi!!!”
No	“@DrSJaishankar What a foolish tweet. EM is squabbling with minister of opposition just to help hisboss चौथी पास राजा to win karnataka elections without caring about damage he is doing to India reputation.”
Comment	“Congress today released another list of seven candidates for the upcoming Karnataka Assembly Elections. On Saturday, Jagdish Shettar tendered hisresignation and joined Congress a day later.”

3.4. Data Classification

The augmented data was labeled manually into three distinct categories. Tweets labeled "positive"had a positive sentiment towards the BJP, while those labeled "negative" had a positive sentiment towards the opposition party, Congress. News reports, unbiased questions and other neutral tweets were classified in the "comment" label, due to their lack of support or opposition to anyone side. Sample tweets for each label are given in Table 1.

Although multiple people labeled the data, the reviewers were found have sufficient agreement as per the Cohen's Kappa metric. The score, calculated using the equation 1, gave a k value of 0.813, indicating "substantial agreement". The metric was calculated using a set of common tweets that all reviewers were made to assess.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

3.5. Data Analysis

An exploration of the augmented data was performed using WordCloud, a library designed to find the most commonly used words and phrases in a piece of text. Three separate Wordclouds were created, one for each data label, namely "positive", "negative", and "comment". This can be seen in Figure 1. When analyzing the most common words for each sentiment, a few neutral words such as "people", "election", and "vote" are seen in both Wordclouds. However, certain terms such as "Indian" are more commonly used by the pro-BJP tweets, while terms such as "politics" are more common among the pro-Congress tweets. Other common rhetoric in the pro-Congress tweets include references to the state Congress party leader Siddaramaiah, words mocking the BJP Prime Minister Narendra Modi, and the term "Hindutva", a right wing ideology sometimes negatively associated with the BJP. Pro-BJP tweets included many references to Dr. S. Jaishankar, India's foreign minister, and the word "stranded", both of which were trending during an online debate regarding Siddaramaih's response to news regarding Indians stranded in Sudan in April. They also contained many campaign slogans for BJP, such as "BJP4India", in support of the party's bid in the national elections in 2024
.

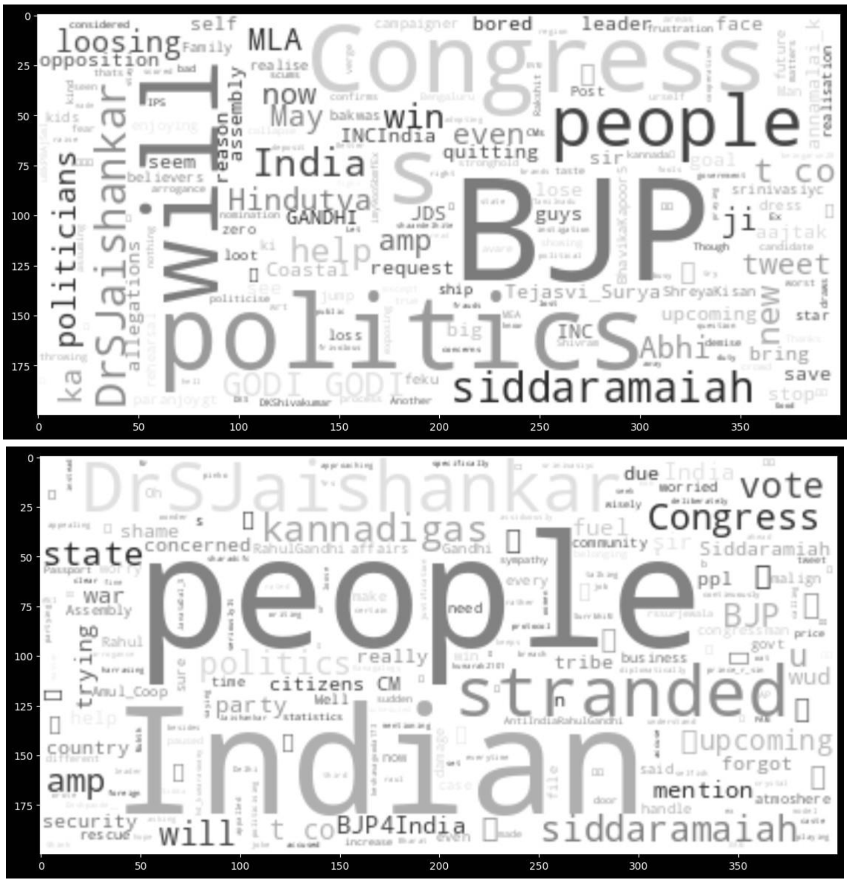

Figure 1. The Wordclouds for the positive and negative sentiment labels

3.6. Sentiment over Time

The relative sentiment towards both parties in this election over time were also analyzed. A sample of the collected tweets were grouped by day and the ratio of positive to negative tweets for each group was calculated. This was then plotted on a line graph, displaying the change in relative sentiment over time. The ratio of positive (pro-BJP) to negative (pro-Congress) tweets fluctuated at about 0.95, implying a slight pro-Congress bias in the overall data set. A few notable exceptions were found, however. The ratio increased to 1.37 on during the time between April 18 to April 20; this corresponded to anti-Congress sentiment online following Siddaramiah's tweet mentioned previously. Other such fluctuations occurred to a lesser extent preceding the election, following which the ratio decreased to 0.89; this could be expected given the landslide victory of Congress in the election.

4. METHODOLOGY

In this section, we provide a detailed account of the rigorous methodology employed in our study, encompassing the utilization of transformer-based models, hyperparameter tuning, and evaluation metrics, following the comprehensive analysis and pre-processing of the collected data.

4.1. Transformers

To delve into sentiment analysis in the context of the Karnataka elections, we leveraged cutting-edge transformer-based models, specifically designed for sentiment analysis in Indic languages. Notably, we utilized BERT (Bidirectional Encoder Representations from Transformers) [1] and IndicBERT [2], which have demonstrated exceptional proficiency in capturing semantic nuances and contextual information in textual data. Figure 2 illustrates the architecture of one of the top-performing transformer-based models used in our study.

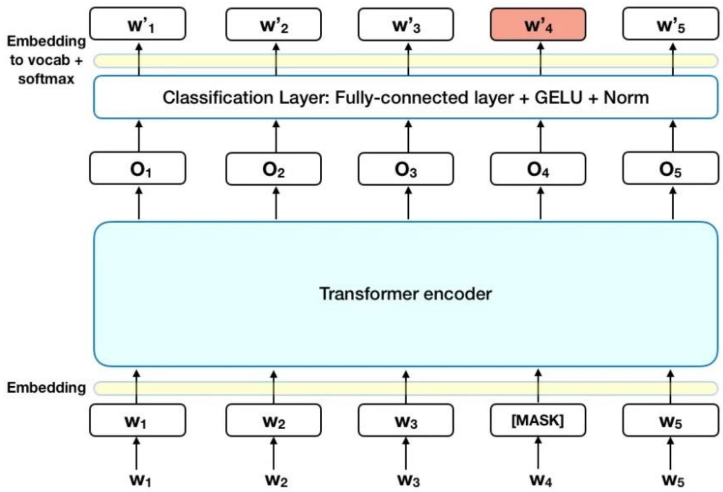

Figure 2. The architecture of BERT, a bidirectional transformer used in the study.

4.2. Models

For this study, BERT served as the primary transformer. It underwent pre-training on Next Sentence Prediction and Masked Language Modeling, which endowed it with groundbreaking bidirectional capabilities [6]. This unique pre-training strategy allowed BERT to predict masked words in random inputs while simultaneously acquiring higher- order distributional statistics.

Additionally, we harnessed the power of IndicBERT as the primary multilingual transformer. Built upon BERT's architecture and training data, IndicBERT encompassed over 11 Indic languages, including Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu [8]. Among these, Kannada and Hindi predominated in the tweets used for the study, although a few Telugu and Tamil tweets were also present.

RoBERTa played a significant role in fine-tuning the models. It employed dynamic masking instead of BERT's static masking. Its training data consisted of BookCorpus, housing 11,038 books, English Wikipedia, over 124 million tweets, and tens of millions of articles on CC- News. Furthermore, we incorporated XLNET, which was trained using BookCorpus, English Wikipedia, Giga5, ClueWeb, and Common Crawl, resulting in a combined 158 gigabytes of training data. Its algorithm rivaled that of BERT across numerous metrics.

DistilBERT, trained on data similar to BERT, exhibited an algorithm running 60% faster than its counterpart, while maintaining over 95% accuracy. Its efficiency allowed for a more robust and efficient training process.

DeBERTA, trained on English Wikipedia, BookCorpus, OPENWEBTEXT, content from Reddit, and a subset of CommonCrawl (STORIES), outperformed both BERT and RoBERTAthrough the implementation of two novel techniques – a disentangled attention mechanism and an enhanced mask decoder. The former encoded words and computed their attention weights using two vectors in relation to content and relative position, while the latter utilized absolute positions in the decoding layer to assist in pre-training [23].

4.3. Hyperparameter Tuning

Hyperparameter tuning is a pivotal step in the optimization of transformer-based models for sentiment analysis, requiring a profound understanding of their intricate interactions and effects on model performance. In our study, we delved into a comprehensive exploration of various hyperparameters, revealing their distinct influences on the efficacy of the models.

The learning rate, a critical hyperparameter, governs the step size during gradient descent, impacting the speed and stability of model convergence. A judicious selection of the learning rate within the range of [1e-5, 1e-3] was crucial. Higher learning rates accelerated convergence but risked overshooting the optimal parameters, leading to suboptimal solutions. Conversely, lower learning rates ensured steady progress but might slow down training considerably, potentially hindering model optimization.

The batch size significantly affected the optimization dynamics and memory utilization during training. We experimented with batch sizes ranging from 8 to 64. Larger batch sizes expedited training by processing more samples in parallel, reducing computation time. However, they also consumed more memory, making it challenging to train on resource- constrained devices. Smaller batch sizes allowed more frequent parameter updates, potentially improving model generalization, but the increased frequency could come at the cost of prolonged training times.

The number of layers and hidden units directly influenced the model's capacity to capture complex patterns in sentiment expression. We explored layer sizes of 64, 128, and 256, along with hidden units in the range of 64 to 512. Deeper models with a higher number of hidden units exhibited enhanced expressive power, potentially leading to superior performance. However, excessively deep models risked overfitting, particularly when training data was limited. On the other hand, smaller models might lack the representational capacity to fully grasp intricate sentiment nuances.

The number of attention heads played a crucial role in determining the models' ability to capture interdependencies between words in the input text. We experimented with 4 to 16 attention heads. Increasing the number of attention heads allowed for more fine-grained analysis of semantic relationships, resulting in models that better understood contextual dependencies. Nevertheless, an elevated number of attention heads also introduced additional computational overhead, making it essential to strike a balance between performance gains and computational cost.

The dropout rate was instrumental in regularizing the model during training to prevent overfitting. Employing dropout during training inhibited neurons from becoming overlyreliant on specific features, ensuring better generalization. We tuned the dropout rate to optimize this regularization effect, avoiding both underfitting and excessive regularization that could hinder the model's capacity to learn complex sentiment patterns.

The number of training epochs represented a critical factor in determining when model training reached a suitable convergence point. We experimented with varying numbers of epochs, ranging from 5 to 50. Too few epochs might lead to underfitting, where the models fail to capture intricate patterns in the data. Conversely, too many epochs risked overfitting, causing the models to memorize the training data, leading to poor generalization on unseen data.

By employing advanced hyperparameter tuning techniques, such as grid search and random search, we navigated the hyperparameter space effectively. The fine-tuned transformer-based models, combined with our diverse and representative dataset, allowed us to achieve remarkable accuracy in capturing nuanced sentiment dynamics during the Karnataka elections. In the subsequent section, we will present comprehensive results, with a particular focus on IndicBERT's efficacy in sentiment analysis.

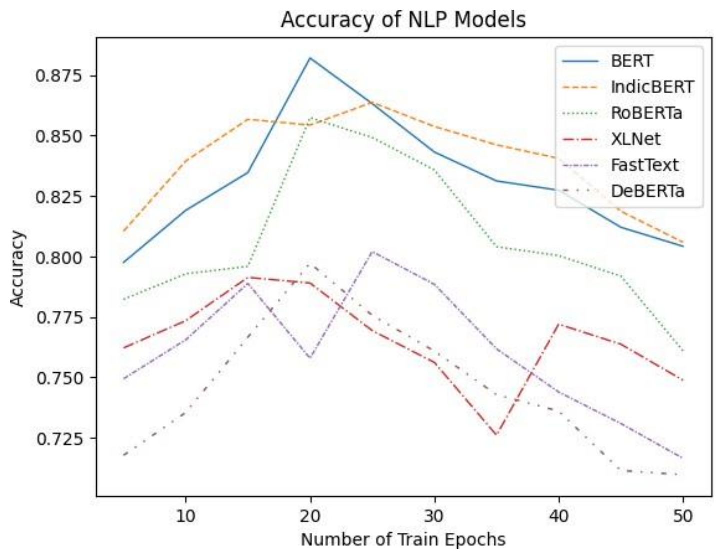

Figure 3. A comparison of training epochs to the accuracy of the models used in the study.

4.4. Evaluation Metrics

Precision, recall, and F1 score. Accuracy measured the overall correctness of sentiment predictions, while precision and recall provided insights into the models' ability to accurately identify positive and negative sentiments. The F1 score, a harmonic mean of precision and recall, provided a balanced assessment of the models' performance.

To address any bias stemming from an uneven ratio of opinionated to neutral tweets in the training data, we utilized the data augmentation techniques mentioned in Section 3. These techniques enhanced the diversity and representativeness of the training data, leading to more robust and unbiased models.

By employing these sophisticated evaluation metrics, we quantitatively assessed the accuracyand effectiveness of our sentiment analysis models, providing meaningful insights into the sentiment dynamics surrounding the Karnataka elections. The detailed results of our analysis, with a particular focus on IndicBERT's performance, will be presented and discussed in the subsequent sections.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the experimental results and analyze the accuracy achieved in predicting the sentiment towards the Bharatiya Janata Party's (BJP) defeat or the Indian National Congress's (INC) victory in the Karnataka elections.

5.1. Results

We evaluated the performance of our sentiment analysis models on the dataset using the evaluation metrics described earlier. Table 2 summarizes the accuracy achieved in predicting sentiment towards the BJP's defeat or the INC's victory.

The results demonstrate the effectiveness of the transformer-based models, particularly IndicBERT, in accurately predicting sentiment in the Karnataka election tweets. Although BERT had a higher accuracy than IndicBERT, the better F1 score of the latter implies that it was the best-performing model. This could be expected due to this model being specifically trained on 12 Indian languages, including Kannada and Hindi. Regardless, all models leveraged their ability to capture contextual information and linguistic nuances, enabling them to decipher the sentiment expressed in Indic languages with high accuracy.

Table 2. Accuracy of all sentiment classification models

Model	**Accuracy**	**F1 Score**
IndicBERT	0.87	0.92
BERT	0.88	0.87
RoBERTa	0.86	0.84
XLNet	0.80	0.83
FastText	0.80	0.74
DeBERTa	0.80	0.80
Baseline	0.36	0

5.2. Analysis

In-depth analysis of the sentiment analysis results revealed interesting insights into the public sentiment towards the BJP's defeat or the INC's victory in the Karnataka elections. The sentiment analysis models successfully captured the prevailing sentiment trends, providing valuable information on the public perception of the political landscape. This analysis helps in understanding the underlying factors contributing to sentiment patterns and can assist political analysts and decision-makers in gaining insights into public sentiment dynamics.

Overall, the experimental results validate the efficacy of our methodology in sentiment analysis for the Karnataka elections, with transformer-based models achieving high accuracy in predictingsentiment towards the BJP's defeat or the INC's victory.

5.3. Prediction of electoral margins

The models' sentiment analysis were also applied to determine the outcome of the 2023 Karnataka elections. The fine-tuned models were given a set of 100 random tweets, and labeled them with the categories of "positive", "negative", and "comment". The resulting ratio of positive and negative tweets was then used to determine the predicted outcome of the election. The ratios do not include the neutral tweets; thus, they do not add to 100. The results are given below in Table 3. It is noteworthy that all models unanimously predicted Congress as the victor in the elections, which was true in the actual elections. The ratio of Congress to BJP, given in the table as well, is also of interest, as it shows how most models labeled a larger volume of tweets as supporting Congress.

Table 3. Election Outcome Predictions

Model	Ratio	Prediction
IndicBERT	62:18	Congress
BERT	60:20	Congress
RoBERTa	75:15	Congress
XLNet	72:19	Congress
FastText	63:27	Congress
DeBERTa	57:22	Congress
Baseline	65:24	Congress

5.4. Implications and Relevance to Indian Politics

The findings of our study have significant implications for political decision-making in the context of sentiment analysis of social media data. By employing various NLP models and data augmentation techniques, we were able to accurately predict the sentiment towards the defeat of the BJP or the victory of the INC in online discussions. These findings can provide valuable insights for Indian political parties, policymakers, and campaign strategists, especially preceding the 2024 Lok Sabha elections.

5.4.1. Improved Sentiment Analysis

The accurate prediction of sentiment towards political events can help political parties gauge public opinion, understand voter sentiment, and tailor their campaign strategies accordingly. Our study demonstrates that multilingual NLP models, such as IndicBERT, can effectively analyze sentiment in social media data related to political events in countries where English may not be the primary language of communication. By leveraging these models, political parties in such nations can gain a deeper understanding of public sentiment and adapt their messaging and outreach strategies to align with the prevailing sentiment.

5.4.2. Identifying Key Issues and Policy Prioritization

Analyzing sentiment data can help political parties identify key issues and prioritize policyinitiatives based on public sentiment. As seen with the timeline of sentiment, social media is a way for political entities to see real-time reactions to their policies and actions. By tracking sentiment towards specific policy areas or events, parties can align their agendas with the concerns and aspirations of the public faster than ever. Furthermore, sentiment analysis can aid in identifying larger sentiment drivers, such as economic issues, social justice, or governance, whichcan guide policy formulation and communication strategies in the long term.

5.4.3.Targeted Campaign Strategies

Our study also highlights the potential of data augmentation techniques in improving the accuracy of sentiment analysis models. By incorporating techniques such as random swap and data translation, we were able to augment the training data and enhance the performance of the models. This augmentation can lead to more robust sentiment analysis models, enabling political parties to design targeted campaign strategies based on nuanced sentiment analysis.

Overall, the findings of our study provide valuable insights into sentiment analysis in the context of political decision-making and campaigning in India. The use of NLP models and data augmentation techniques can significantly improve sentiment prediction and aid political parties in understanding public sentiment, shaping campaign strategies, and responding effectively to emerging trends.

6. CONCLUSION AND FUTURE RESEARCH

In conclusion, the findings of this study hold significant implications for political decision-making in the context of Indian politics. The successful application of sentiment analysis using transformer-based models and data augmentation techniques offers valuable insights into public sentiment during elections. By accurately capturing and understanding sentiment dynamics, political actors can devise informed strategies, identify voter concerns, and tailor their messaging to resonate with prevailing sentiments.

The utilization of transformer-based models, including BERT, IndicBERT, RoBERTa, XLNET, DistilBERT, and DeBERTA, has proven to be instrumental in analyzing sentiment in Indic languages, particularly in the case of the Karnataka elections. These state-of-the-art models excel in capturing semantic nuances and contextual information, allowing for a deeper understanding of sentiment expressions within the diverse linguistic landscape of Karnataka.

Moreover, the employment of data augmentation techniques further enhanced the representativeness and generalizability of the dataset. The augmentation methods effectively addressed bias concerns and ensured a balanced representation of opinionated and neutral tweets, contributing to the models' robustness and accuracy.

As the field of AI and natural language processing continues to evolve, transformer-based models and data augmentation techniques will likely play an increasingly vital role in political sentiment analysis. Their adaptability to diverse languages and ability to capture complex patterns make them indispensable tools for understanding public sentiment in multilingual and culturally diverse societies like India.

6.1. Future Research

While our study contributes to the understanding of sentiment analysis in political contexts, there are several avenues for future research. Firstly, exploring more advanced NLP models and architectures could potentially yield even better results in sentiment analysis. Additionally, investigating the application of sentiment analysis to other political events, such as key debates or speeches, or domains could provide a broader understanding of public sentiment and political discourse. For example, a potential source of Congress' victory could have been the Bharat Jodo Yatra of Congress leader Rahul Gandhi, which spanned multiple months prior to the election. Analysis of sentiment regarding the Yatra could be collected and analyzed similar to the WordCloud or sentiment timeline presented in this paper.

Furthermore, extending the analysis to targeted multilingual sentiment analysis, could address the unique challenges and nuances of sentiment prediction in diverse linguistic contexts. For example, the state of Karnataka has at least 3 well defined dialects of Kannada, in addition to over 150 local dialects. A Kannada-specific analysis of this election's outcome could be performed as well, using districts or dialect borders to divide the state into different regions for targeted analysis of results and sentiment. Incorporating contextual information, such as geographic location or demographic factors, even within the state, could also enhance the accuracy and granularity of sentiment analysis in the future.

A limitation of our work, stemming the cultural atmosphere of India, is potential bias in the user base of Indian twitter. The average age of Twitter users in the nation is lower than the national average, implying that the younger generation uses the platform more. Other online platforms, such as Facebook and Whatsapp, contain more data from the older generations, and would likely provide better insight into the political sentiments of a larger group of Indians. Failure to acknowledge this limitation could lead to potential issues for political entities as they may not be able to effectively reach all of their electoral base.

In conclusion, sentiment analysis using NLP models and data augmentation techniques offers valuable insights for political decision-making. By leveraging these techniques, political parties and policymakers can gain a deeper understanding of public sentiment, tailor their strategies, and effectively respond to emerging trends. Future research in this area has the potential to further refine sentiment analysis methods and contribute to the development of robust tools for political analysis and decision-making.

REFERENCES

[1] Cambria, E., & Hussain, A. (2012). Applications of Sentiment Analysis. In the Handbook of Natural Language Processing (pp. 733-756). Springer.

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[3] Gohil, N. K., & Vasanwala, S. (2018). Twitter sentiment analysis for Indian politics. In 2018 3rd International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.

[4] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter:What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

[5] Mandal, A., Srinivasan, A., Gupta, R., & Ray, C. (2020). Sentiment Analysis in Indian Languages: A Comprehensive Overview. In 2020 International Conference on Natural Language Processing and Computational Linguistics (NLPCL) (pp. 1-6). IEEE.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171-4186).

[7] Akhtar, M. S., Yadav, P., & Rehman, M. A. (2020). A Comparative Study of Multilingual Models for Sentiment Analysis in Indic Languages. In 2020 International Conference on Inventive Research in Computing Applications (pp. 1-5). IEEE.
[8] Joshi, M., D'Souza, J., Bhattacharyya, P., & Carman, M. J. (2020). IndicBERT: A Language Model forIndian Languages. arXiv preprint arXiv:2005.00813.
[9] Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party won the German election of 2009or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". Social Science Computer Review, 30(2), 229-234.
[10] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.
[11] Wang, W., & Manning, C. D. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2 (pp. 90-94).
[12] Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 1320-1326).
[13] Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data. Springer Science & Business Media.
[14] Aggarwal, C. C. (2018). Neural networks and deep learning. Springer.
[15] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751).
[16] Nguyen, D., & Shirai, K. (2015). Phrases in sentiment analysis: A systematic literature review. Journalof Natural Language Processing, 22(1), 151-192.
[17] Mishra, S., Mishra, R., & Yadav, M. (2019). Sentiment Analysis on Twitter Data Using Machine Learning Algorithms. International Journal of Emerging Technologies and Innovative Research,6(7), 1-6.
[18] Wei, Jason & Zou, Kai. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performanceon Text Classification Tasks.
[19] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-Level Convolutional Networks for Text Classification. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (Montreal, Canada) (NIPS'15). MIT Press, Cambridge, MA, USA, 649–657.

AUTHORS

Pranav Gunhal is an Artificial Intelligence Researcher affiliated with the Artificial Intelligence Coalition. His research interests include applications of AI and political science. His past work, which focused on transformer architectures as well as political sentiment analysis, has been published in accredited journalssuch as IEEE.