Multilingual Indian Political Posts Analysis: Predicting Topical Stance and Political Leaning using BERT

Nitin Bisht, Samman Sarkar

[a]*Netaji Subhas University of Technology*

Abstract

In contemporary Indian politics, social media initiatives have become an important tool for political parties, which are widely used for publicity and electioneering In elections, platforms like Twitter play an important role in politics advocacy and in the development of parties and candidates. This study examines the effectiveness of various strategies using social media in predicting election results. A model for predicting the outcome of India's 2024 general elections is being developed through machine learning, specifically sentiment analysis. Voters are increasingly turning to social media to express their views and participate in political discourse, making understanding and analyzing the dynamics of these platforms necessary to assess electoral behavior. Using longitudinal methods, this study aims to shed light on the potential impact of social media sentiment on electoral outcomes in the Indian context. We also conduct image post analysis by fine-tuning the idefics -9B model to extract relevant features from political images, which will then be integrated with textual data for comprehensive political stance and party affiliation classification using BERT.

Keywords:
Deep learning, Rapid Miner, Pattern recognition, Twitter, Intelligent applications, Sentiment analysis, Machine learning

1. Introduction

In the digital age, social media channels have emerged as influential spaces to shape political discourse, shape public opinion and influence election results The advent of networks such as Twitter and Reddit has made information dissemination possible democracy, and allow users to discuss a wide range of topics including politics in a virtual battleground Political parties compete for attention and support. Understanding the complex nuances of Indian politics on social media is paramount for political analysts, policy makers and researchers. Sentiment analysis techniques have become important tools for measuring public sentiment and predicting electoral outcomes. Nowadays, there has been tremendous growth in the use of social media platforms to express thoughts, feelings,

or opinions on topics. Sensitivity analysis methods are generally classified as dictionary, machine learning, and hybrid methods.

Recent years have seen a rise in research focused on evaluating computer sentiment votes, increasing use of social media data mining insights and vote prediction Sentiment analysis includes artificial intelligence systems where individuals share their sentiments beliefs or opinions on a range of topics including political events and identity To do so , social media platforms are increasingly being used, and political parties are launching campaigns, leading to wider discussion and expression of views revealed. It can be hard to separate personal opinions from trends on social media. But platforms like Twitter have become preferred sites for data collection due to their structure, allowing users to express themselves in 280 characters or less Several studies show Twitter's ability to gather information and analyze sentiment in which researchers use sentiment from tweets, big data and developed models for sentiment analysis In this context, the in-

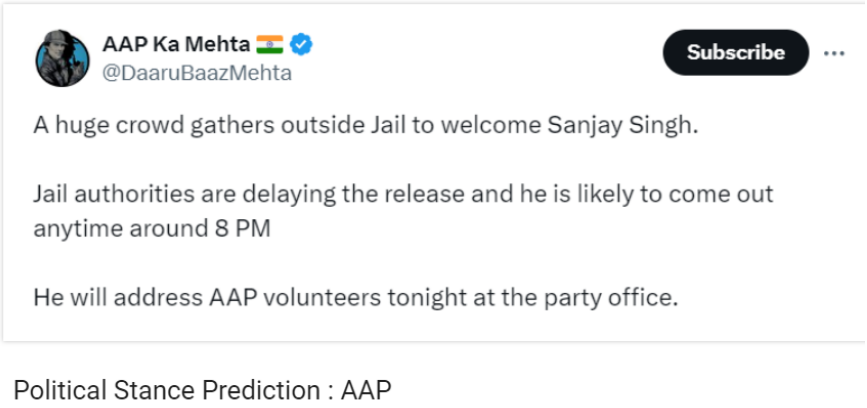

Figure 1: Political Stance Prediction based on tweet

troduction of advanced natural language processing has transformed perceptual analysis. Some such approaches are the BERT (Bidirectional Encoder Representations from Transformers), IndicBert and XLM-RoBERTa model. BERT, a state-of-the-art language translation model, is popular for its ability to understand context and meaning in textual data.IndicBert and XLM-RoBERTa are the adaptation of Bert specially designed for Indian language The use of BERT allows researchers to capture the subtle nuances and linguistic complexities of social media language, making sentiment analysis methods more accurate and effective. Analysis of the impact of critical events on election forecasting using social media has seen a significant contribution to the research environment in terms of exploring event-based sentiment analysis methods for election forecasting

In this context, our research seeks to contribute to this growing field by focusing on the 2024 general elections and delving into the multilingual nature of Indian politics on Twitter and Reddit.

- We conducted an evaluation of the proposed work, which demonstrated superior performance compared to Naive Bayes and Support Vector Machine.
- Deep learning models are employed to propose an event-based sentiment analysis method for predicting elections.
- Using advanced natural language processing techniques, specifically the various BERT models and optimizing mBERT with the final layer of LSTM, we aim to provide a nuanced understanding of the thematic and political context of social media users to shed light on the digital pulse, and commentary, of Indian democracy on governance, electoral systems and public discourse .
- In addition, our research introduces a pioneering approach by incorporating image analysis using the fine-tuned idefix-9B model alongside advanced natural language processing techniques. This integrated approach aims to capture nuanced visual and textual features from political posts, enhancing the accuracy and depth of our analysis regarding topical stance and political leaning prediction.

2. Related Work

2.1. Predicting Election Results from Social Media Data

Predicting election results from social media data, sentiment analysis techniques were used to analyze public perceptions of the United States presidential election in real time. The results show public perceptions of the election campaign [1].

2.2. Use of Twitter in the 2011 Singapore General Election

The 2011 Singapore General Election demonstrated the usefulness of Twitter and social network applications, showing that Twitter is central to election campaigns and voter turnout analysis. Seven political parties focused on it, with a Mean Absolute Error (MAE) of 5.13%. However, a definitive relationship between the percentage of votes against and the number of tweets received could not be established [2].

2.3. Concept Extraction Techniques

A variety of concept extraction techniques have been used, including unigrams, bigrams, part-of-speech tags, etc. Using unstructured data representations as sources of results has shown improvements over previous work [4,5].

2.4. Post-Election Sentiment Analysis in Nigeria

In February 23, 2019, Nigeria held presidential elections. Many Nigerians expressed their views on or criticized various presidential candidates on social media. Post-election sentiment analysis from Nairaland, a targeted social network for Nigerians, was used to identify polarity sentiment (i.e., positive or negative) in 118,421 cases using supervised machine learning and dictionary-based methods [3].

2.5. Addressing Financial Fraud in Instagram Stories

The Chrome extension "ScamSpot," developed in 2023 by Stefan Erben and Andreas Valdis, addresses financial fraud in Instagram stories. This extension aims to combat such fraudulent activities by providing users with a way to detect and report scams in Instagram posts. Overall, the "ScamSpot" Chrome extension reflects the way in which financial fraud is addressed on social media platforms such as Instagram, providing valuable insights into public sentiment towards political parties through the analysis of tweets [6].

2.6. A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election

Studies have previously used Twitter to study communication and discussion in the US. during presidential elections [13,14,15,16,17]. Jacob and so on. [18] conducted a sensitivity analysis of Twitter data collected ten days before and after Election Day. Their aim was to relate the sensitivity of the conversation to the characteristics of Twitter users, such as the number of their followers, the duration of their activity, and the number of tweets Joyce and Deng [19] focused on the relationship between between the sentiment of the tweet and the subject line so the internal hashtag of the tweet. In our research, we are also analyzing tweets collected before and after Election Day However, we relate sentiment analysis of tweets to the status of the tweet and the status of its author. In doing so, we are able to find groups of users who behave differently in a volatile election cycle.

Almuhimedi and so on. [19] first presented a method to quantify and evaluate deleted tweets on the platform, and Zhou et al. [20] further modeled deletion behavior and applied sentiment analysis to deleted tweets. Meeks [21] used the method to analyze deleted tweets of politicians to show how political campaigns use a strategy to hide and extract information from voters Our study is a new addition to this area of research, as we used deleted tweets from one of the Twitter We will cover all the observed sentiments about the main political event. Excluding deleted or inaccessible tweets from the sentiment analysis would provide an incomplete picture of the circulation of opinions over the election cycle.

3. Model Considerations

In a study that focused on analyzing multilingual Indian political terminology to explore subjective ideologies and political ideologies using BERT, it is necessary to select appropriate samples to obtain the effective and accurate. Simple models such as naive, SVM are also used while IndicBERT and XLM-RoBERTa are also used.

- **Naive Bayes Classifier:** Naive Bayes classifier is a simple probabilistic model based on Bayes' theorem. Despite its simplicity, Naive Bayes is adept at information segmentation tasks and performs relatively well, making it suitable for large data sets. It assumes independence of parts from each other, which may not always be true in real-world situations, but its ease of implementation and quick training make it a valuable starting point comparable to text classification tasks although not as good as M-BERT It provides with a useful measure.

- **Support Vector Machines (SVMs)**: Support Vector Machines (SVMs) are a conceptually similar supervised algorithm for classification and regression. SVMs attempts to set up a separation hyperplane that maximizes the distance between the classes in feature space. SVMs is tested on withstanding high dimension input data and the importance in the area of text segmentation tasks. In terms of studying SVMs combined with M-BERT, it is easy to model a tradeoff between complexity of the model, computation time, and prediction accuracy by comparing multiple text vectorizers. Since SVMs allows the non-linear function relationship between different components, it is an option for working with the case of a multilingual M-BERT model in political discourses.

- **BERT (Bidirectional Encoder Representations from Transformers):** The transformer-based model BERT was used to create a model that transforms natural language processing tasks. BERT captures the bidirectional flow of context from sequences, which enables him to understand the structure of language and complex relationships. However, when it extends to many languages like Indian, it becomes Multilingual BERT which is more designed for multilingual political discourse analysis in India. Subtle and under rational understanding contextual relationships can identify contextual rather than clear, well-defined relationships, Mass appropriate ideological ideology and political development.

- **IndicBERT (Indic Bidirectional Encoder Representations from Transformers):** IndicBERT is an adaptation of BERT specially designed for Indian languages. It retains the structure and capabilities of BERT but is trained on different types of Indian language data. IndicBERT enables efficiency and comprehensibility of text in Hindi, Bengali, Tamil, etc. making it particularly suitable for analyzing political terms and ideas in the Indian context By using IndicBERT researchers can ensure that

the model is prepared to address current linguistic and cultural nuances in Indian politics.

- **XLM-RoBERTa (Cross-lingual Language Model RoBERTa):** XLM-RoBERTa is a version of RoBERTa trained on multilingual data including Indian languages. It extends the framework of RoBERTa to handle multilingual tasks, capable of understanding and providing information in multiple languages without prior language-specific training XLM-RoBERTa has particular advantages for tasks involving multilingual information, and justifies an examination of multilingual political discourse in India. Its ability to capture contextual relations between languages enhances its effectiveness in predicting political development and ideological development in different languages.
- **Optimizing Multilingual BERT with LSTM for Enhanced Natural Language Processing Across Languages:** Optimizing Long Short-Term Memory (LSTM) Optimizing Long Short-Term Memory (LSTM) on Multilingual BERT (mBERT) includes improving the overall performance of language understanding fashions throughout numerous languages. By fine-tuning LSTM architectures with mBERT embeddings, researchers purpose to improve sequential studying and contextual comprehension. This synergy leverages mBERT's multilingual illustration abilties and LSTM's potential to seize long-range dependencies. Through meticulous parameter tuning and training strategies, consisting of gradient clipping and gaining knowledge of charge scheduling, practitioners decorate version convergence and generalization. The optimized LSTM on mBERT not handiest fosters move-lingual transfer mastering however also advances numerous herbal language processing duties, contributing to the improvement of sturdy and versatile language models.

4. Methodology

The methodology used in this paper outlines the systematic process by which political postings on social media were classified. Using machine learning methods and advanced natural language processing (NLP) models, our approach aims to describe the subject representation and political representation of various textual multimedia objects. The following sections describe our data collection methods, preprocessing steps, sample selection, analysis metrics, and future plans for extending our analysis to multimedia sources Figure 1 outlines the methodology followed in our research for analyzing political posts on social media platforms. It illustrates the step-by-step process starting from data collection and preprocessing, followed by feature extraction using advanced NLP techniques such as BERT, IndicBert, and XLM-RoBERTa.

4.1. Data Collection and Annotation

A research project to capture the multidimensional content of Indian politics on Twitter, especially during the 2024 general elections in India, embarked on

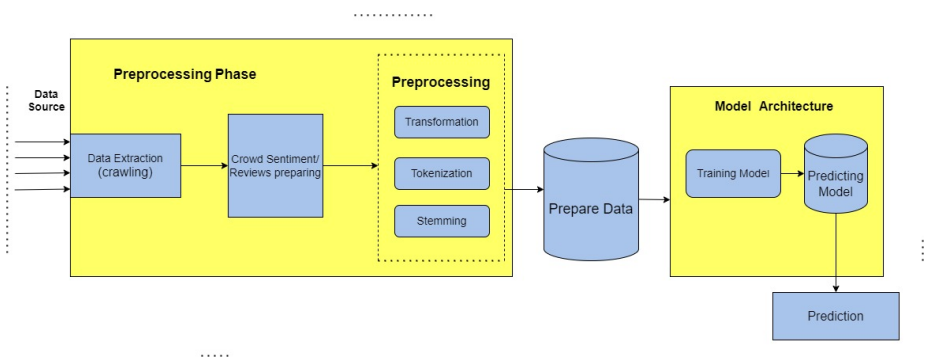

Figure 2: Methodology

an extensive data collection and documentation journey. Twitter's advanced search capabilities played a key role in filtering tweets in multiple politically relevant languages, including English, Hindi, Bengali, to ensure that a wide range of viewpoints were specially monitored and monitored Hashtags related to popular politicians like Modi and Kejriwal, and captured the conversations and sentiments surrounding prominent politicians. The inclusion of such hashtags enriched the dataset with discussions and opinions about influential politicians, and contributed to a broader search using advanced keyword identification techniques monitored tweets containing specific political terms and topics. Keywords related to major political parties like BJP, Congress, AAP, TMC, SP etc. were carefully maintained to ensure that the dataset contained a wide range of political affiliations and ideologies Such a strategy this increased the relevance and topic relevance dataset by anticipating the topical status of party affiliations In addition to the relevant subsequent research tasks, the project adopted a dynamic focus on capturing political meme hashtags trending like PoliticalMemes and ElectionHumor.

Table 1 presents our sample posts of selected curated political posts dataset, shows various issues related to political parties Like BJP, Congress, AAP, TMC and SP. The dataset spans wide Written and visual content, including tweets, memes, posters and news stories, The articles, reflect the multifaceted nature of political issues on social media platforms.

Post Text	Party
Modi ji ke liye vote karein! Modi2024	BJP
Kejriwal should be released IStandWithKejriwal	AAP
Priyanka Gandhi to hold roadshow in Varanasi on May 15.	Congress

Table 1: Sample posts of political posts dataset

Figure 2 visually displays a representative tweet from our dataset alongside its predicted political stance.

These hashtags not only added some humor and satire to the dataset but also

provided insight into how political discourse on social media platforms shaped memes and humorous comments about political figures and groups was systematically, and reflected public sentiment and engagement of key political events and identity By targeting these figures Data collection became very beautiful. This targeted approach ensured the inclusion of multiple perspectives and micro analyses, contributing to a well-rounded dataset for subsequent analysis and classification work Systematic descriptions were made and each post was labeled according to political correct group support It also embraced data, advanced search capabilities, keyword discovery channels , hashtag tracking , and the use of targeted content curation , Twitter has created a powerful corpus for analyzing Indian politics in-depth analysis, especially in the context of discussions about top politicians like Modi and Kejriwal in anticipation of the 2024 general elections.

Figure 3 visually shows the distribution of rows or instances for each political party in our data set. By showing the number of users or information about parties like BJP, Congress, AAP, TMC, and SP, this statistic also gives insight into each party's participation and representation of our dataset. Understanding the distribution of information across political parties is important to ensure balanced training and evaluation of machine learning models, and will ultimately make our political position prediction and sentiment analysis projects more accurate and unbiased.

Figure 4 presents a visual representation of the distribution of rows or instances across different languages in our dataset. By quantifying the number of posts or instances in languages such as English, Hindi and Bengali, this figure provides insights into the linguistic diversity of political discourse on social media platforms. Understanding the prevalence of different languages in our dataset is crucial for ensuring balanced training and evaluation of machine learning models across linguistic variations.

Table 2 presents a structured overview of the number of tweets associated with each political party, categorized by different languages in our dataset. By grouping tweets based on language categories such as English, Hindi, Bengali, and others, this table provides valuable insights into the linguistic distribution of political discourse surrounding various parties. Analyzing the count of tweets across languages helps in understanding the linguistic preferences of users when discussing different political entities on social media platforms. This analysis aids in identifying language-specific engagement levels, sentiment patterns, and thematic focus related to each political party, contributing to a comprehensive understanding of multilingual political discourse dynamics.

Figure 5, presented as a bubble chart with the y-axis representing political parties and the x-axis representing languages, further delves into the distribution of rows or instances across languages in our dataset, building upon the insights provided in Figure 4. By highlighting the specific counts of posts or instances for each language and political party, including any outliers or dominant languages, this figure offers a detailed perspective on the linguistic and party-wise composition of our dataset.

By categorizing tweets based on language categories such as English, Hindi,

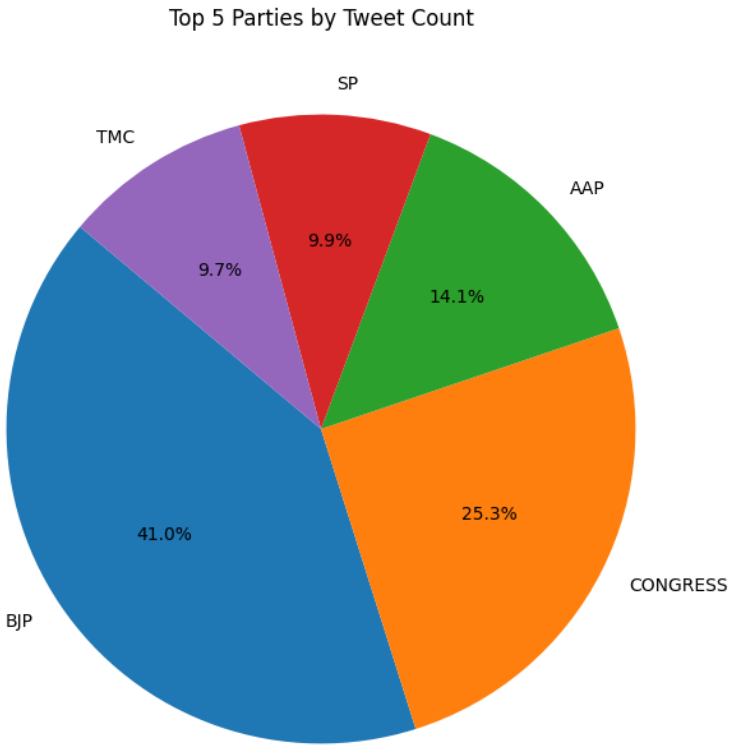

Figure 3: No. of rows for each party

Party	English	Hindi	Bengali
AAP	85	312	35
BJP	1183	39	28
CONGRESS	511	93	163
SP	59	109	131
TMC	113	64	116

Table 2: Number of tweets for each party grouped by language.

Bengali, and others, we gain valuable insights into the linguistic distribution of political discourse surrounding various parties. These insights enhance our understanding of multilingual political dynamics, contributing to more robust and nuanced analyses.

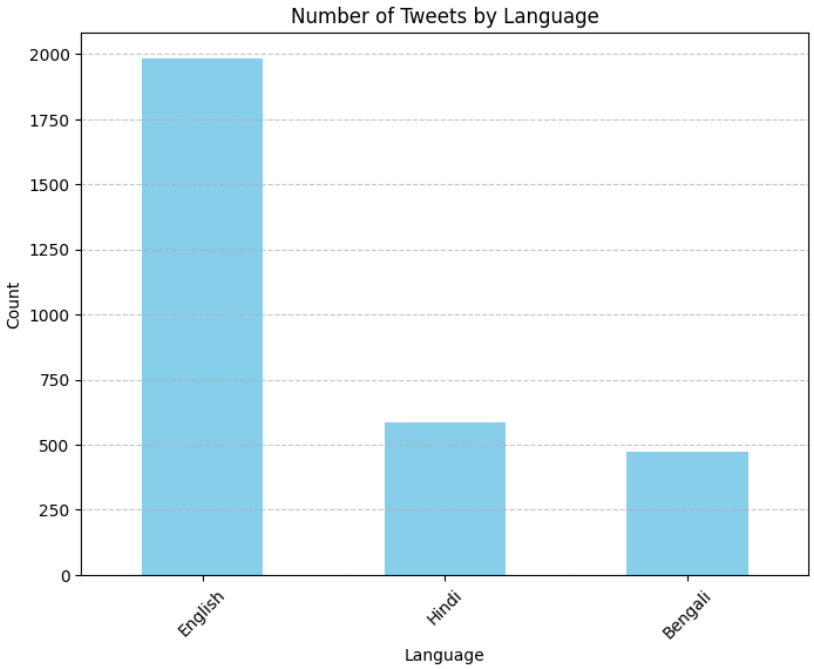

Figure 4: No. of rows for each language

4.2. Data Preprocessing

After data collection, a careful data preprocessing phase was applied to ensure the quality and integrity of the dataset This involves removing redundant characters, emojis and URLs, followed by tokenization made the script for later analysis. The dataset was then formatted in a tabular format with two main columns: one tweet text and the other referring to the corresponding political parties Political parties represented in the dataset included BJP, Congress, AAP, TMC and SP. This categorization served as an important basis for subsequent research work, including sentiment analysis, topic status prediction, and political classification Not annotating each tweet with its associated political party enabled researchers to search political nuances in the data set.

The preprocessing pipeline for image information includes several steps enabled by the "Clean()" function. First, this function removes punctuation, numbered words, and special symbols from all descriptions. Additionally, it converts all descriptions to lowercase for consistency. Then, tokenization is applied to the data set using a fixed dictionary size of 8,464. For pre-drawing, it is important to prepare the images before bringing them into the drawing. This involves resizing each image to the appropriate resolution for the chosen model architecture, such as (299*299) for Xception or (224*224) for VGG16 and then flattening the images and scaling the pixel values, a process known as normalization, to ensure uniformity and facilitate model meeting After this

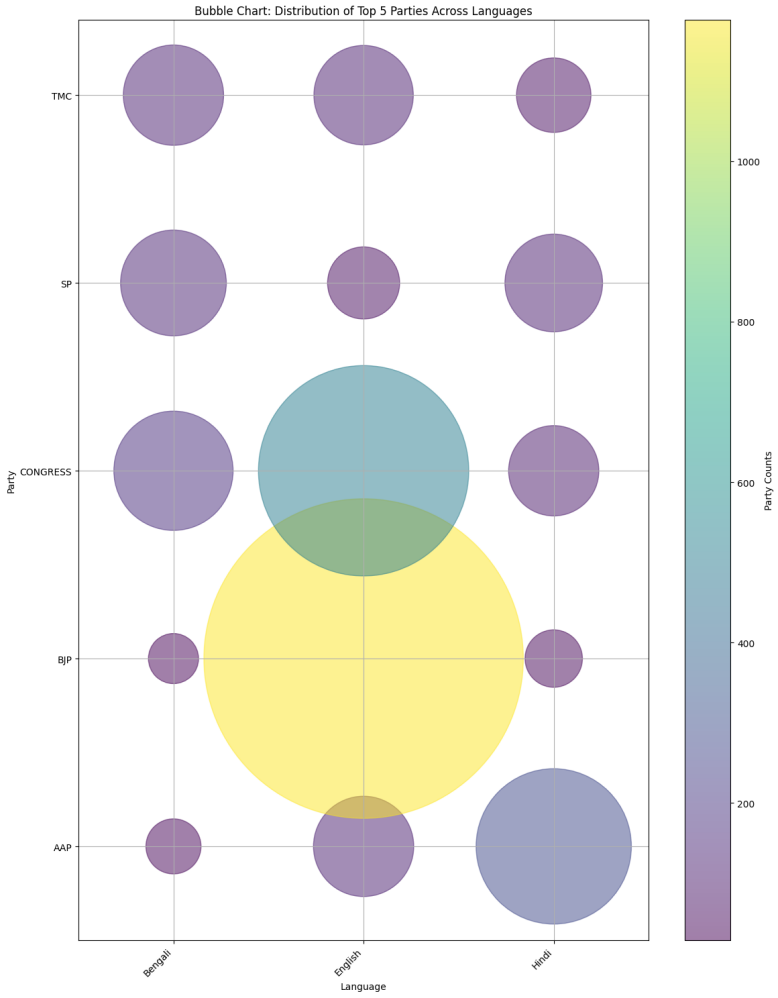

Figure 5: Distribution of tweets across parties and languages

preprocessing step, image descriptions and data are formatted accordingly for subsequent model training and analysis.

4.3. Encoding

Numerous efficient learning models necessitate digital vector inputs rather than textual data. Therefore, we utilized the bag-of-words technique to transform text into digital vectors, incorporating word scores and conducting feature extraction. Our research selects optimal classification algorithms based on thorough literature reviews, ensuring the use of effective and suitable models for our dataset. Subsequently, our model is constructed in accordance with these findings, prioritizing accuracy and efficiency in classification tasks. the encoding

processes necessary to generate the data sets used in many of our political research programs.

4.4. Data Split

The dataset, totaling 3041 instances, was meticulously divided into three distinct subsets to facilitate robust model training, validation, and evaluation. The training dataset, comprising 80% of the total data size (2432 instances), was used exclusively to train machine learning models, allowing them to learn intricate patterns and features. The validation dataset, constituting 20% of the total data size (609 instances), played a crucial role in fine-tuning hyperparameters, detecting overfitting, and assessing model performance before final evaluation. A separate testing dataset, not explicitly mentioned, was kept aside for the final evaluation, ensuring that the models' generalization to unseen data and real-world performance could be accurately assessed. Careful attention was given to maintaining class distribution across subsets to prevent bias and ensure representative learning and testing scenarios. The label mapping used in this study includes the following categories: {'BJP': 0, 'CONGRESS': 1, 'AAP': 2, 'TMC': 3, 'SP': 4}, which corresponds to the political parties under consideration for classification tasks.

4.5. Model Evaluation and Testing.

In evaluating the performance between SVM, Naive Bayes, Multilingual BERT, IndicBert and a XLM-Roberta model based on Recall, F1 Score, Accuracy, and confusion matrix metrics, we followed a rigorous methodology to ensure fair comparison and reliable results.

Firstly, we split our dataset into training, validation, and test sets, maintaining a balanced distribution of samples across different classes and languages. We trained each model using the training set and fine-tuned hyper parameters using the validation set to optimize performance.

Next, we evaluated the models on the test set to assess their real-world performance. For each model, we calculated Recall, F1 Score, and Accuracy metrics to measure their ability to correctly classify tweets and Reddit posts based on their topical stance and political leaning.

The confusion matrix provided a detailed breakdown of the model's predictions, showing the true positives, false positives, true negatives, and false negatives for each class. This analysis helped us understand the strengths and weaknesses of each model in terms of correctly identifying relevant instances and avoiding miss classifications.

By systematically comparing these metrics across SVM, Naive Bayes, Multilingual Bert, XLM-RoBERTa and IndicBERT models, we gained insights into their respective capabilities in handling multilingual Indian political text data and predicting topical stance and political leaning with high accuracy and robustness. These results are crucial for informing decision-makers and researchers about the most effective model for such tasks in real-world applications.

4.6. Integrating LSTM with mBERT for Improved Accuracy

Adding the LSTM layer on top of the mBERT model has proven to be a significant improvement, showing a remarkable accuracy increase up to 94.58%. This integration leverages the contextual understanding of mBERT languages and the sequential learning capabilities of LSTM. Leveraging mBERT's pretrained context adjustments as input features for LSTM levels, the model gains deeper linguistic perspective and improves its predictive capabilities. This approach not only increases accuracy but shows the ability to combine pretrained models with different architectures to achieve significant growth in natural language processing tasks.

5. Experimental Results

Table 3 presents a sample of posts from the dataset along with their corresponding labels denoting political parties or affiliations. Table 2 summarizes the performance metrics of the 5 models: SVM (Support Vector Machine), Naive Bayes, Multilingual BERT, IndicBERT and XLM-RoBERTa evaluated on the test set.

Accuracy is a metric used to measure the overall correctness of a classification model. It calculates the ratio of correctly predicted instances to the total instances in the dataset. The formula for accuracy is given by:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalInstances}$$

Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify relevant instances among all relevant instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives. The formula for recall is given by:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It considers both false positives and false negatives and is particularly useful when classes are imbalanced. The formula for F1 score is given by:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Training the mBERT model was achieved using a series of steps to enhance its performance and accuracy to classify political content on social media. The training was conducted through a structured approach whereby the training loss and validation metrics were constantly compared to estimate the model's performance.

The training loss for the mBERT model was evaluated using a standard cross-entropy loss function, which measures the discrepancy between predicted probabilities and actual labels across training epochs. The cross-entropy loss

loss aids the optimization of the model by ensuring that the learning process enables it to optimize the parameters to a point where the prediction errors are minimized.

During training, we utilized a specific number of epochs, with epoch 4 being a crucial checkpoint in our training process. By epoch 4, the model has completed enough iterations to capture meaningful patterns and relationships in the training data without overfitting. such a compromise between the complexity of the model and its basing on data is an important criterion for stable work and high quality on new data.

Validation was performed regularly during training to assess the model's performance on a separate validation dataset, ensuring that it generalizes well to new, unseen examples. The validation loss and accuracy are the criteria by which we evaluate the quality of further work of the model on political text while avoiding overfitting.

In our training process, the validation loss was measured at 0.1177, indicating the average loss on the validation dataset per sample. A lower validation loss suggests that the model's predictions align well with actual labels, reflecting higher accuracy and generalization.

Figure 6 visually represents the training loss trajectory of the mBERT model across training epochs. The plot showcases how the model's training loss decreases over successive epochs, indicating improved learning and convergence towards optimal parameter values. Monitoring training loss helps us track the model's progress during training, ensuring that it learns meaningful representations from the data and improves its predictive capabilities. The downward trend in training loss observed in Figure 6 affirms the effectiveness of our training process and the model's capacity to capture intricate patterns in political text data for accurate stance prediction.

Training accuracy and validation accuracy are crucial metrics in evaluating the performance of machine learning models, including the mBERT model used in our research for political stance prediction.

During the training process, training accuracy measures the proportion of correctly classified instances within the training dataset. It reflects how well the model learns from the training data and adjusts its parameters to make accurate predictions on seen examples. On the other hand, validation accuracy assesses the model's performance on a separate validation dataset, providing insights into its ability to generalize to new, unseen examples and avoid overfitting.

In our training process, achieving a validation accuracy of 0.9392 signifies that the mBERT model correctly classified approximately 93.92% of instances in the validation dataset. This high validation accuracy indicates the model's effectiveness in accurately predicting political stances across diverse textual data related to different political parties or ideologies.

Figure 7 visualizes the training accuracy trajectory of the mBERT model across training epochs. The plot illustrates how the model's training accuracy improves over successive epochs, reflecting its ability to learn and generalize from the training data. Monitoring training accuracy provides insights into the model's learning progress and convergence towards optimal performance.

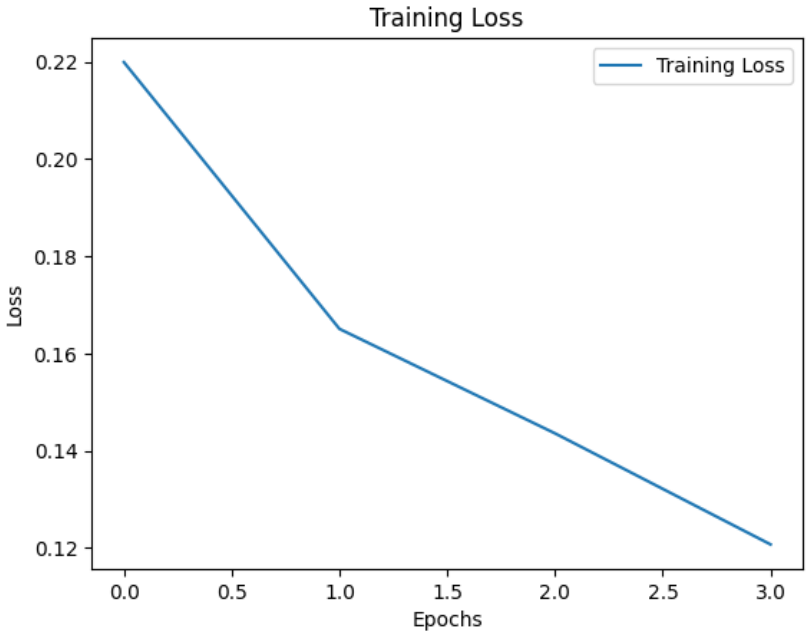

Figure 6: Training Loss for mBERT

The upward trend in training accuracy observed in Figure 7 corresponds to the model's improved ability to correctly classify instances within the training dataset as training progresses. This improvement in training accuracy aligns with the model's enhanced capability to learn intricate patterns and nuances in political text data, contributing to higher validation accuracy and robust performance on unseen political content.

Figure 8 presents the heatmap visualization representing the classification results of the Naive Bayes model. The heatmap showcases the model's accuracy and performance concerning both political affiliations and topical stances, providing insights into its strengths and areas for improvement.

Figure 9 displays the heatmap for the Support Vector Machine (SVM) model, offering a detailed view of its classification accuracy across various political affiliations and topical stances. The heatmap visualization helps in understanding how well the SVM model distinguishes between different classes and topics.

Figure 10 illustrates the heatmap visualization for the Multilingual BERT model, a state-of-the-art language model known for its contextual understanding. The heatmap highlights the model's performance nuances across different political affiliations and topical stances, showcasing its ability to capture intricate patterns and contexts.

Figure 11 showcases the heatmap for the Indic BERT model, designed specifically for languages with Indic scripts. The heatmap provides insights into the model's classification accuracy concerning political affiliations and topical

Figure 7: Training Accuracy for MBERT

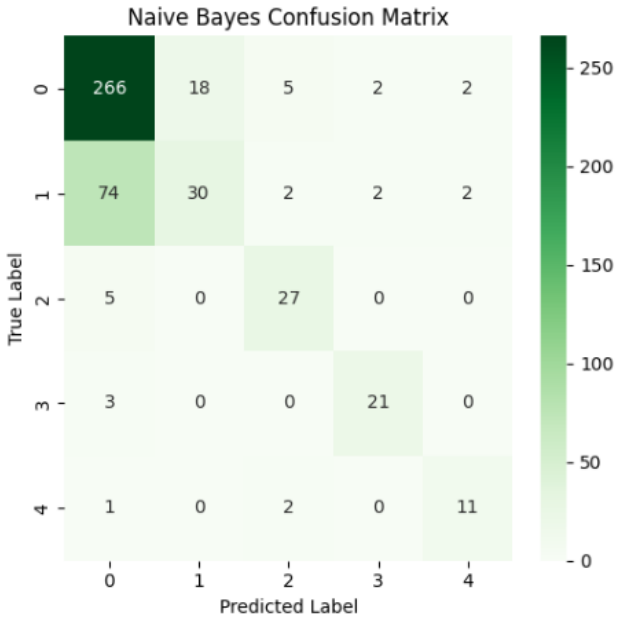

Figure 8: Heat-map of Naive Bayes

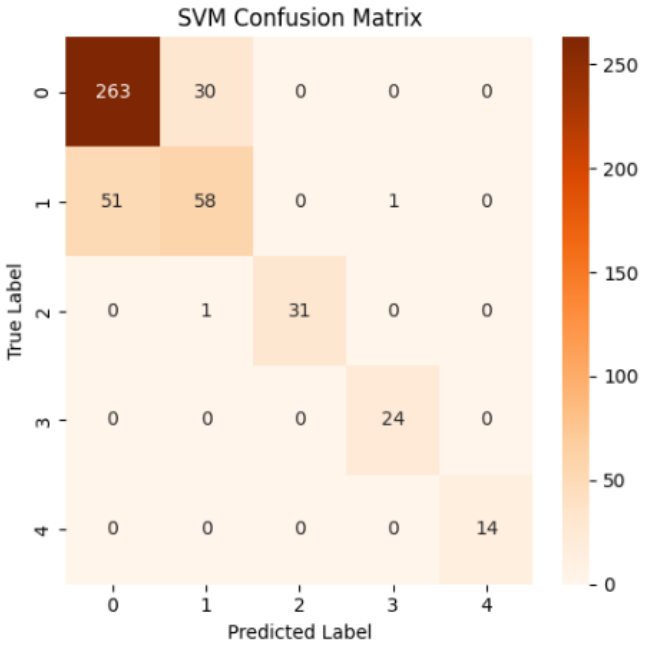

Figure 9: Heat-map of SVM

stances, particularly in the context of languages using Indic scripts.

Figure 12 represents the heatmap visualization for the XLM-RoBERTa model, a cross-lingual language model known for its multilingual capabilities. The heatmap highlights the model's accuracy and performance across various political affiliations and topical stances, showcasing its effectiveness in handling diverse languages and contexts.

Figure 13 represents the heatmap visualization for the optimzed model of multilingual BERT(mBERT) using LSTM layer on it, a cross-lingual language understanding model, leveraging multilingual embeddings and sequential learning. The heatmap highlights the model's accuracy and performance across various political affiliations and topical stances, showcasing its effectiveness in handling diverse languages and contexts.

These heatmaps offer a comprehensive visual representation of each model's classification performance, aiding in comparative analysis and understanding the models' strengths and weaknesses in handling different classification tasks within the political and topical domains.

One remarkable aspect of our study was the inclusion of different languages in the dataset, which yielded satisfactory results across the board. Despite the inherent challenges of multilingual data processing, our models demonstrated robustness and accuracy in classifying content across diverse linguistic contexts. This success underscores the versatility and adaptability of our approach, showcasing its effectiveness in handling linguistic variations and nuances present in

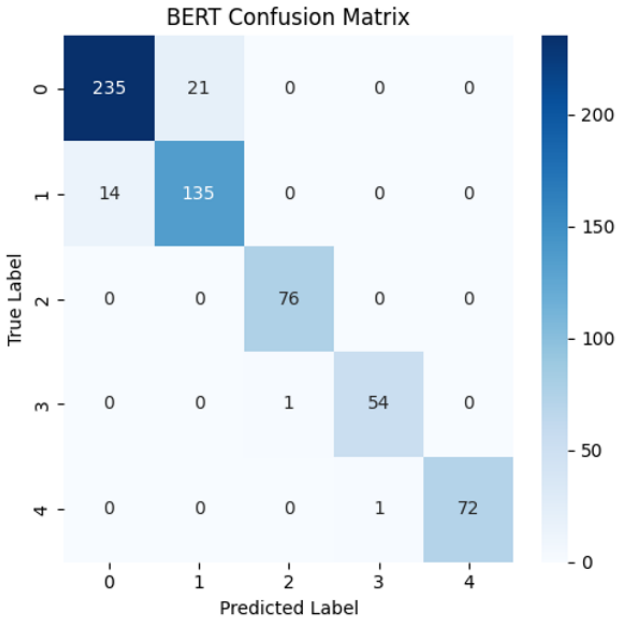

Figure 10: Heat-map of mBERT

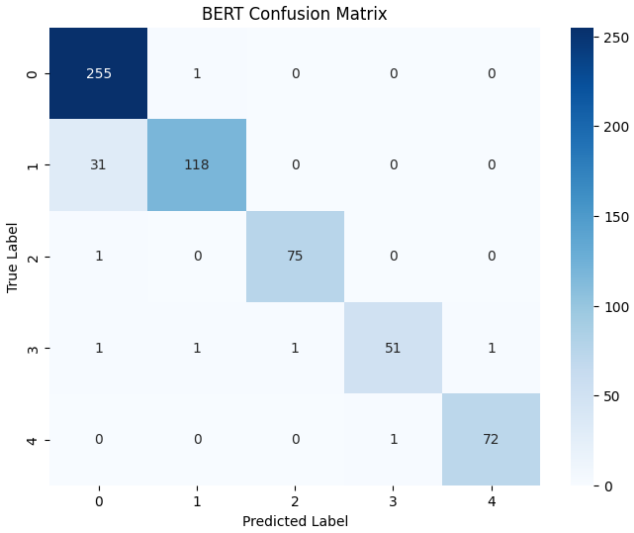

Figure 11: Heat-map of IndicBERT

multilingual Indian political content. the successful handling of multilingualism added significant depth and breadth to our analysis, contributing to a more comprehensive understanding of political discourse in the Indian context in context

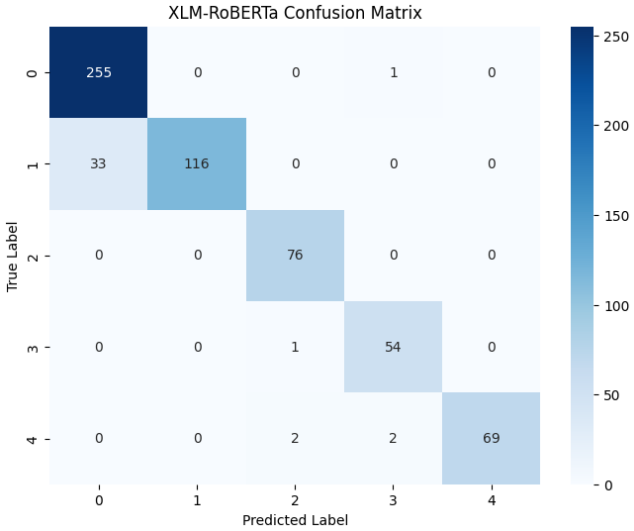

Figure 12: Heat-map of XLM-RoBERTa

of the upcoming general elections.

Table 3 presents a comprehensive comparison of five different models based on their accuracy, recall, and F1-score metrics in classifying political content. The models evaluated include Naive Bayes, Support Vector Machine (SVM), mBERT, IndicBERT, and XLM-RoBERTa, each representing distinct approaches to natural language processing and machine learning.

- Naive Bayes: The Naive Bayes model achieved an accuracy of 0.7505, recall of 0.7505, and F1-score of 0.7213. It shows moderate performance across metrics, but it falls short compared to more sophisticated models.
- SVM: The SVM model outperformed Naive Bayes with an accuracy of 0.8245, recall of 0.8245, and F1-score of 0.8182. Its higher accuracy and balanced recall and F1-score shows improved performance in classifying political stances compared to Naive Bayes.
- mBERT: The mBERT model demonstrated superior performance with an accuracy of 0.9392, recall of 0.9392, and F1-score of 0.9395. Its high accuracy and balanced recall and F1-score highlight its effectiveness in capturing nuanced linguistic features and political sentiments.
- IndicBERT: IndicBERT closely rivals mBERT with an accuracy of 0.9376, recall of 0.9376, and F1-score of 0.9362. Its performance showcases the adaptability of language models specifically designed for Indian languages in political stance prediction tasks.
- XLM-RoBERTa: The XLM-RoBERTa model also demonstrates strong performance with an accuracy of 0.9360, recall of 0.9360, and F1-score

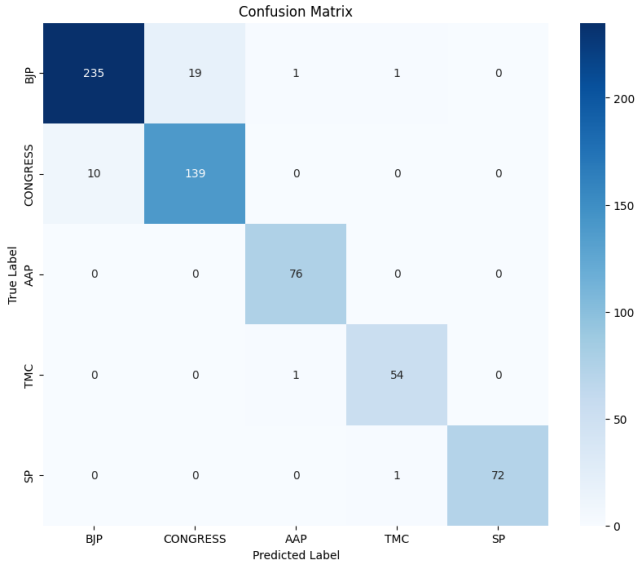

Figure 13: Heat-map of mBERT+LSTM

of 0.9342. Its competitive metrics emphasize the efficacy of cross-lingual models in analyzing multilingual political content.

- mBERT+LSTM: The combined mBERT and LSTM model achieved outstanding results, boasting an accuracy of 94.58%, recall of 94.58%, and an impressive F1-score of 94.65%. This significant improvement over the baseline mBERT model, which had an accuracy, recall, and F1-score of 93.92%, showcases the synergistic power of integrating LSTM architecture with mBERT's multilingual contextual embeddings.

Among the observed models, mBERT performed significantly better in terms of high accuracy, recall, and F1-score, while Naive Bayes exhibited comparatively low performance in all metrics

Figure 14 visually presents the comparison based on the accuracy metrics of the five models. Bar graphs or plots enable to visualize the detailed performance patterns among the models, and confirm the findings from Table 3. MBERT, IndicBERT, XLM-RoBERTa better performance than traditional models such as Naive Bayes and SVM highlight the importance of a there is a need for advanced natural Language processing and language models previously trained in political situation prediction and sentiment analysis tasks.

6. Image Post Analysis

We took on the task of fine-tuning the Idefics-9B model for political image analysis, with the aim of increasing the depth of our overall analysis. We

Model	Accuracy	Recall	F1-Score
Naive Bayes	0.7505	0.7505	0.7213
SVM	0.8245	0.8245	0.8182
MBERT	0.9392	0.9392	0.9395
IndicBERT	0.9376	0.9376	0.9362
XLM-RoBERTa	0.9360	0.9360	0.9342
MBERT+LSTM	0.9458	0.9458	0.9465

Table 3: Comparison of different models

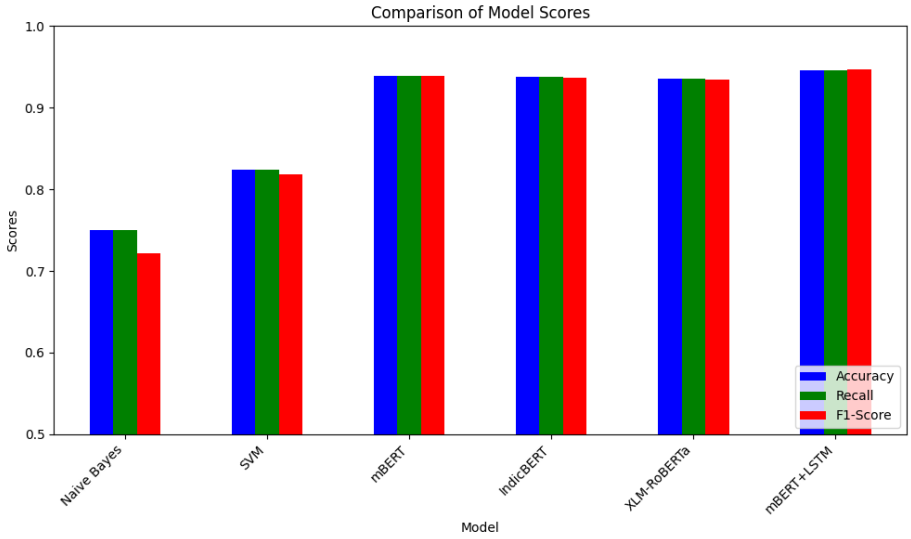

Figure 14: Comparison of different models

created a separate database of political picture words representing parties like BJP, Congress, AAP, TMC and SP, which were carefully labelled for training. We then fine tuned the Idefics 9B model, which is a open-access visual language model to extract key information from various visual sources, including memes and posters, such as identifying relevant individuals,and understanding relevance signals, and distinguishing political themes. The results of the image analysis process were seamlessly integrated into the BERT model, providing us with a nuanced and classification of subjective context and political ideologies based on textual and visual cues in social media posts.

In Figure 15, we provide an example of an image featuring Prime Minister Narendra Modi, along with accompanying text. The output of this figure demonstrates the model's prediction, accurately identifying the politician in the image and extracting relevant text associated with the image. This integration of image analysis with text processing showcases the model's capability to understand both visual and textual elements, contributing to enhanced accuracy

in identifying political figures and analyzing associated content on social media platforms.

Meanwhile, Figure 16 depicts our image analysis methodology, focusing on the fine-tuning process of the idefics-9B model. The output from this fine-tuned model is then used as input in the BERT model, enabling us to leverage both visual and textual features for more accurate political stance prediction and sentiment analysis across multimedia content on social media platforms.

This integration allowed us to leverage both visual and textual information, enhancing the accuracy and depth of our political stance prediction and sentiment analysis across multimedia content on social media platforms.

Figure 15: Identifying context and text in image

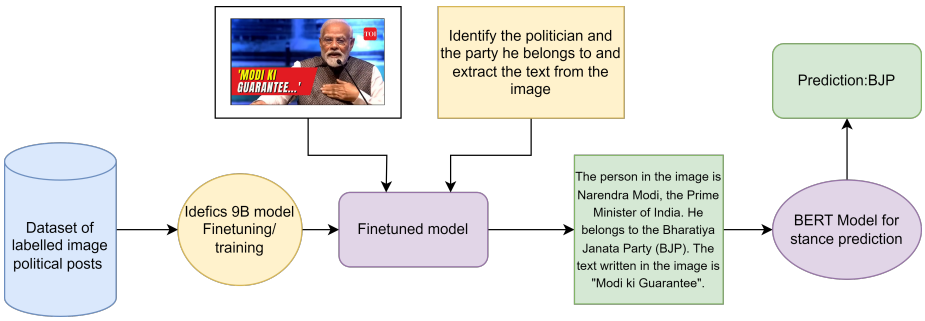

Figure 16: Image analysis methodology

7. Future Work : Chrome Browser Extension

In terms of future work, we plan to expand our research by implementing a user-centric Chrome browser extension integrated with a REST API that utilizes a fine-tuned BERT model for automatic detection and classification

of the political stance in tweets across various social media platforms. This extension will not only extract tweet text seamlessly from the user's browsing environment but will also provide users with intuitive options to choose between modes that either highlight tweets based on political stance or filter out tweets from specific political affiliations such as BJP, Congress, AAP, TMC, and SP.

Furthermore, we aim to enhance user engagement and participation by incorporating a dynamic system within the extension, allowing users to report any misclassified tweets. This feedback mechanism will be instrumental in continuously refining and improving the classification model, thereby ensuring more accurate and reliable results over time. By bridging cutting-edge machine learning techniques with user-friendly interfaces, our future work seeks to empower users with deeper insights into the nuanced political discourse on social media platforms, fostering informed and meaningful interactions within the digital political landscape.

8. Conclusion

In this paper, we have discussed the study to the relationship between social media and the transformation of political discussion and sentiment analysis particularly in the context of the 2024 general elections in India. We show the skillful utilization of state of the art natural language processing techniques with BERT, IndicBert, XLM-Roberta and optimized BERT+LSTM. Through these techniques, we understand more deeply the multi-language world and the discussion on Twitter and Reddit using tools developed and configure with the datasets used. Our matrix shows the decent performance deviation learning methods compared with general methods -Naive Bayes and Support Vector Machine. Our research process includes rigorous testing and the use of key performance metrics such as Recall, F1 Score, Accuracy, confusion matrices etc. to assess each method's potential subjective and political contextualization.

Moreover, our novel approach to the integrated exploitation of image analysis based on the idefics-9B model fine-tuned for our purposes and textual analysis has expanded our analysis by capturing nuanced visual and textual features from political posts . This comprehensive methodology has substantially improved the precision and completeness of our analysis on the topic stance detection and political preference anticipation, thus offering a precise picture of the digital heartbeat and commentary of the Indian democracy. As a result, our investigation makes a valuable contribution to the field of computational social science and points to the further potential for research in the domain of multimedia data analysis and real-time monitoring of social media interactions, thus empowering stakeholders with actionable ideas for decision making and government in the digital era.

References

[1] Miftahul Qorib, Rahel S Gizaw, Junwhan Kim, "Impact of Sentiment Analysis for the 2020 U.S. Presidential Election on Social Media Data", ICMLT

'23: Proceedings of the 2023 8th International Conference on Machine Learning Technologies.

[2] M. Skoric, N. Poor, P. Achananuparp, E.-P. Lim, and J. Jiang, "Tweets and votes: A study of the 2011 singapore general election." pp.2583-2591

[3] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis." pp. 347-354.

[4] Hussain, A., M. Ahmad, and I.A. Mughal. Automatic Disease Detection in Wheat Crop using Convolution Neural Network. in The 4thInternational Conference on Next Generation Computing. 2018

[5] B. Pang, and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." p. 271.

[6] Stefan Erben, Andreas Waldis,"ScamSpot: Fighting Financial Fraud in Instagram Comments".

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL), pages 4171–4186. Association for Computational Linguistics.

[8] Beta Priyoko and Ainul Yaqin. 2019. Implementation of Naive Bayes Algorithm for Spam Comments Classification on Instagram. In 2019 International Conference on Information and Communications Technology (ICOIACT), pages 508–513.

[9] CHRISTOPHER IFEANYI EKE , AZAH ANIR NORMAN , AND LIYANA SHUIB "Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model".

[10] Rukhma Qasim, Waqas Haider Bangyal , Mohammed A. Alqarni, and Abdulwahab Ali Almazroi "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification".

[11] Lara Grimminger and Roman Klinger "Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection".

[12] Taraneh Ghandi, Hamidreza Pourreza, Hamidreza Mahyar "Deep Learning Approaches on Image Captioning".

[13] Grover P, Kar AK, Dwivedi YK, Janssen M. Polarization and acculturation in us election 2016 outcomes-can twitter analytics predict changes in voting preferences. Technol Forecast Soc Change. 2019;145:438–60.

[14] Bovet A, Makse HA. Influence of fake news in twitter during the 2016 us presidential election. Nat Commun. 2019;10(1):1–14.

[15] Enli G. Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. Eur J Commun. 2017;32(1):50–61.

[16] Abilov A, Hua Y, Matatov H, Amir O, Naaman M. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. arXiv preprint arXiv:2101.08210 2021.

[17] Yaqub U, Chun SA, Atluri V, Vaidya J. Sentiment based analysis of tweets during the us presidential elections. In: Proceedings of the 18th Annual International Conference on Digital Government Research, 2017; pp. 1–10.

[18] Almuhimedi H, Wilson S, Liu B, Sadeh N, Acquisti A. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, 2013; pp. 897–908.

[19] Zhou L, Wang W, Chen K. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 603–612.

[20] Meeks L. Tweeted, deleted: theoretical, methodological, and ethical considerations for examining politicians' deleted tweets. Informat Commun Soc. 2018;21(1):1–13.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[22] Li Xu, Jun Zeng and Shi Chen, "yasuo at HASOC2020: Fine-tune XML-RoBERTa for Hate Speech Identification"

[23] Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, Praneesh Sharma ,"ML AI IIITRanchi@DravidianLangTech: Fine-Tuning of IndicBERT for Exploring Language-Specific Features for Sentiment Classification in Code-Mixed Dravidian Language"

[24] Introducing IDEFICS: An Open-Access Multimodal AI Model Advancing Transparency

[25] Danae Sánchez Villegas, Daniel Preo¸tiuc-Pietro, Nikolaos Aletras,"Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks"

[26] Mastering BERT: A Comprehensive Guide from Beginner to Advanced in Natural Language Processing (NLP)

[27] A Brief Introduction to BERT by Adrian Tam on January 6, 2023