Multimodal Multilingual Social Media Analysis on Political Leaders

Ankur Tiwari, Anshuman Uniyal, Hrishikesh Khambete, Karanbir Singh, Shashank G. Sharma and Vani Mittal

***Abstract*— The project explores the social media discourse surrounding the Indian general elections, explicitly focusing on opinions on the Sharechat platform. By employing a multi-step approach involving data scraping, text extraction, sentiment analysis, and scoring mechanisms, the project aims to explain political contenders' sentiments and perceptions better. Through developing a user-friendly interface, users can engage in exploratory data analysis, gaining valuable insights into the minute landscape of political sentiment on social media.**

I. INTRODUCTION

In the era of social media, understanding public sentiment towards various topics, personalities, or events is essential for businesses, policymakers, and researchers alike. ShareChat, as a popular social media platform, hosts a vast array of user-generated content spanning multiple languages and diverse topics. Analyzing sentiment within this ecosystem presents unique challenges due to the content's multilingual nature and prevalence of image-based posts. This project addresses these challenges by developing a comprehensive sentiment analysis framework tailored to ShareChat. The project aims to extract, translate, and analyze sentiments expressed in ShareChat posts by leveraging a combination of text and image processing techniques. The ultimate goal is to provide stakeholders with actionable insights into public sentiment trends, facilitating informed decision-making and strategic planning.

Through a systematic approach involving data acquisition, text extraction, translation, sentiment analysis, and user interface development, this project seeks to offer a robust and user-friendly solution for sentiment analysis on ShareChat. By evaluating and comparing different sentiment analysis models, including state-of-the-art techniques, the project aims to identify the most effective approach for capturing sentiments accurately in this unique social media landscape.

This report provides an overview of the project's methodology, including the tools and techniques employed at each stage. We also present the evaluation results, highlighting the performance of various sentiment analysis models and discussing their implications. Finally, we offer insights into the project's contributions to sentiment analysis research.

II. PROBLEM STATEMENT

The central challenge of this project revolves around utilizing the dataset extracted from ShareChat. On this social platform, users express diverse opinions and emotions, particularly towards influential political figures such as Narendra Modi, Rahul Gandhi, Arvind Kejriwal, and many other prominent leaders. Our primary objective is twofold: first, to develop robust methodologies that seamlessly convert image-based content into multilingual textual data, and second, to leverage an advanced sentiment analysis model capable of translating sentiments expressed across diverse languages and evaluating a score, thereby retrieving the emotional expressions expressed by users.

III. MOTIVATION

1) Insights into Public Sentiment: The project provides significant insights into public sentiment on social media platforms, offering a nuanced understanding of perceptions and attitudes towards political figures. Through comprehensive sentiment analysis, it illuminates the prevailing sentiments expressed by users, thus enriching our understanding of digital discourse dynamics.
2) Validation of Traditional Polling Methods: Through comparative analysis between sentiment analysis results and traditional polling methods, the project validates the efficacy of sentiment analysis as a complementary approach for measuring public opinion. By corroborating findings from both methodologies, it underscores the value of sentiment analysis in capturing the zeitgeist of public sentiment in real-time.
3) Academic Research and Knowledge Generation: Serving as a valuable resource for academic researchers, the project facilitates in-depth exploration of public opinion, social media dynamics, and sentiment analysis methodologies. By sharing datasets, methodologies, and insights, it fosters collaboration and knowledge exchange within the academic community, thereby driving further research and innovation in sentiment analysis and social science research methods.

IV. LITERATURE SURVEY

A. Topic 1 : Multilingual Text Classification and Sentiment Analysis

1) Introduction: Multilingual text classification and sentiment analysis have garnered increasing attention due to the proliferation of user-generated content across diverse linguistic landscapes, notably on social media platforms like Twitter. This section provides an overview of recent advancements in this domain, highlighting seminal contributions and avenues for further exploration.
2) State-of-the-Art Techniques: Recent developments have witnessed the adoption of sophisticated techniques, prominently BERT-based approaches, for multilingual text classification and sentiment analysis.

These methodologies, exemplified in studies such as Gauba et al. (Paper 1), underscore the efficacy of leveraging pre-trained language models for cross-lingual tasks. Additionally, comparative analyses, as delineated in the literature, offer insights into the performance and generalizability of multilingual approaches across various sub-tasks.

3) Gauba et al. (Paper 1): Gauba et al. introduce the MARC Dataset and a Multilingual Twitter Dataset, furnishing researchers with resources for training and evaluating models in real-world contexts. The study elucidates the application of BERT-MC for semantically similar text retrieval and underscores the significance of domain-agnostic and multilingual solutions. Despite notable achievements, limitations pertaining to negation handling and the scope of word models warrant consideration, shaping the trajectory of future investigations.
4) Comparison of Sentiment Analysis Techniques: In parallel, comparative evaluations, as exemplified in the literature, shed light on the efficacy of diverse sentiment analysis techniques across multiple languages. Challenges such as translation errors and feature space expansion underscore the intricacies inherent in cross-lingual sentiment analysis, necessitating nuanced methodological considerations.
5) Under-Resourced Languages: The landscape of sentiment analysis extends beyond dominant languages, as evidenced by the systematic review conducted by Paper 3. Deep learning methodologies emerge as pivotal in augmenting sentiment analysis outcomes for under-resourced languages, underscoring the transformative potential of advanced techniques in addressing linguistic diversity.

B. Topic 2 : Dataset Utilization in Sentiment Analysis

1) Introduction: Datasets serve as foundational pillars in the development and evaluation of sentiment analysis models, offering researchers invaluable resources for training and benchmarking purposes. This section delves into the utilization of key datasets in sentiment analysis research, elucidating their significance and associated challenges.
2) Key Datasets Used: Across various studies, a plethora of datasets are employed, each tailored to specific research objectives. From the MARC Dataset and Multilingual Twitter Dataset utilized in Paper 1 to the IIT-Patna Hindi Reviews dataset examined in Paper 4, researchers leverage diverse corpora to explore multifaceted dimensions of sentiment analysis.
3) Analysis Techniques and Dataset Suitability: Methodological choices, as exemplified in Papers 1 and 4, are intricately intertwined with the characteristics of the employed datasets. Aligning analysis techniques with dataset properties is imperative to mitigate biases and enhance model performance, necessitating meticulous scrutiny of dataset suitability.
4) Challenges in Dataset Utilization: Notwithstanding the indispensability of datasets, challenges loom large in their utilization. Issues ranging from bias and domain-specificity to resource availability pose formidable obstacles, necessitating concerted efforts to address these limitations and foster robust research practices.

C. Topic 3 : Aspect-Based Sentiment Analysis

1) Introduction: Aspect-based sentiment analysis (ABSA) offers a granular perspective on sentiment expression, enabling nuanced insights into user opinions across diverse domains. This section delves into recent advancements in ABSA methodologies, elucidating their implications and inherent challenges.
2) Ensemble Models and ABSA: The advent of ensemble models, as demonstrated in Paper 4, heralds a paradigm shift in ABSA research. Leveraging pre-trained mBERT models, researchers circumvent traditional limitations and augment performance through innovative methodologies, paving the way for enhanced sentiment analysis capabilities.
3) Limitations and Challenges: Despite notable advancements, ABSA endeavors encounter multifaceted challenges. Computational exigencies, corpus limitations, and polarity detection intricacies pose formidable hurdles, necessitating concerted efforts to surmount these obstacles and advance the frontiers of ABSA research.
4) Performance and Future Directions: The findings gleaned from Paper 4 underscore the efficacy of BERT-based models in ABSA tasks, while also delineating avenues for future exploration. Constructive synthesis of aspect information and leveraging BERT for sentence pair classification hold promise for addressing existing challenges and propelling ABSA research into new horizons.

Each topic within the research paper is meticulously examined, synthesizing key findings and offering insights into the current state and future trajectory of multilingual text classification, sentiment analysis, and aspect-based sentiment analysis research. Through a comprehensive exploration of seminal contributions and associated challenges, this literature survey contributes to the collective understanding of sentiment analysis methodologies across diverse linguistic landscapes.

V. NOVELTY

This research project presents a novel exploration into the realm of multi-modal and multilingual data analysis, with a specific focus on the ShareChat platform. Despite the exponential growth of user-generated content on social media platforms, ShareChat remains relatively underexplored in academic research. Our endeavor bridges this gap by delving into the complexities of multi-modal data, which encompasses textual, visual, and audio content, thus offering a holistic understanding of user interactions within this unique platform.

Moreover, our research extends beyond conventional monolingual analyses to encompass multilingual data analysis. ShareChat, being a platform that caters to a diverse linguistic user base, provides an ideal milieu for studying the intricacies of multilingual communication. By leveraging advanced computational techniques, we aim to unravel patterns, sentiments, and cultural nuances embedded within the multilingual content shared on ShareChat.

The amalgamation of multi-modal and multilingual analyses in the context of the ShareChat platform represents a significant departure from traditional approaches to social media data analysis. By undertaking this novel endeavor, we aspire to contribute valuable insights into the dynamics of user engagement, content dissemination, and cultural exchange within this burgeoning social media ecosystem. Through rigorous empirical investigation and methodological innovation, our project endeavors to advance the frontiers of both academic research and practical applications in the field of digital communication and social media analytics.

VI. METHODOLOGY

The proposed methodology for analyzing political opinions on social media, mainly focusing on the ShareChat platform, comprises a systematic and multi-step approach to extract meaningful insights from textual data related to various political leaders. Initially, image-to-text conversion is facilitated using Tesseract OCR, a powerful optical character recognition tool meticulously configured to accommodate multilingual content to English translation from image posts. Subsequently, to ensure linguistic uniformity and facilitate comprehensive analysis, the translated text undergoes further processing through DeepTranslator, leveraging its GoogleTranslator module for efficient translation. Following translation, standard text preprocessing techniques are applied, encompassing tasks such as lowercase conversion, punctuation removal, and tokenization, which collectively enhance the quality and consistency of the extracted textual data. The crux of the analysis lies in sentiment analysis, where NLTK's Vader module is employed to assign sentiment scores—positive, negative, and neutral—to each post, providing valuable insights into the prevailing sentiments expressed toward political figures. To aggregate sentiment scores across posts, a normalization step is introduced, scaling composite scores based on post-engagement metrics, thus ensuring comparability and accuracy in sentiment analysis results. Subsequently, aggregate scores are utilized to compute the average sentiment score for prominent political figures like Narendra Modi, Rahul Gandhi, Arvind Kejriwal and many others leaders. This facilitates a comprehensive understanding of public sentiment dynamics surrounding these leaders on the ShareChat platform. This methodology, characterized by its systematic approach and integration of advanced tools and techniques, serves as a robust framework for discerning and analyzing political opinions on social media, contributing to a deeper understanding of the digital discourse landscape.

VII. DATABASE

The database for this project primarily consists of data scraped from the ShareChat platform. This data includes posts, images, and metadata related to political content. The database serves as the raw material for analysis. The data is stored in a structured format, allowing for efficient retrieval and processing. The database is updated regularly with new data from ShareChat to keep the sentiment analysis up-to-date.

The database also stores the sentiment scores generated for each political leader. These scores are calculated based on the ShareChat posts and reflect the public sentiment towards these leaders. The scores are stored in a way that allows for easy retrieval and visualization.

VIII. CODE

The code for this project is written in Python and uses several libraries and tools. The Scarechat Scraper is used to scrape data from the ShareChat platform. Tesseract OCR is utilized for image-to-text conversion, particularly for translating Hindi image posts to English text. The DeepTranslator tool is employed to translate multilingual text into English, explicitly using the GoogleTranslator module from $\text{deep}_t ranslator for eff icient translation.$

The Natural Language Toolkit (NLTK) is used for text preprocessing and sentiment analysis. This includes standard preprocessing steps such as tokenization, stop word removal, and lemmatization. The SentimentIntensityAnalyzer from NLTK is utilized to compute positive, negative, and neutral sentiment scores.

The code also includes a user interface developed to present the sentiment scores visually, allowing users to explore and analyze sentiment trends. This interface is built using web development languages and libraries.

The code is organized in a modular fashion, with separate modules for data scraping, preprocessing, sentiment analysis, and visualization. This modular structure makes the code easy to understand, maintain, and extend.

IX. EVALUATION

1) Ground Truth Labeling: Manual labeling was conducted for each post to establish the ground truth sentiment, with labels assigned as 'positive', 'neutral', or 'negative' based on the intuition of individual labelers. This process ensured the availability of a reliable reference for evaluating the performance of sentiment analysis models.
2) Model Prediction and Labeling: Following ground truth labeling, each post's sentiment score was predicted by the sentiment analysis models. The model assigned the label with the highest sentiment score to each post. This facilitated comparison between the model's predicted sentiment and the ground truth labels.
3) Calculation of Evaluation Metrics: Various evaluation metrics were computed based on the comparison between the actual ground truth labels and the model's

predicted labels. The metrics utilized include Accuracy, Precision, Recall, F1-Score, and ROC-AUC value. These metrics provided comprehensive insights into the performance of each sentiment analysis model.

4) Models Utilized: Multiple sentiment analysis models were employed to calculate sentiment scores, including:
   - distilbert-base-uncased-finetuned-sst-2-english model from
   - Transformers Huggingface
   - Bi-LSTM based Flair model
   - NB-Classifier based TextBlob model
   - VADER model from NLTK
5) Performance Comparison: The performance of each model was compared using the aforementioned evaluation metrics. A state-of-the-art (SOTA) baseline model, the Bi-LSTM based Transformer model, was used as a benchmark for comparison. This facilitated the identification of the most effective sentiment analysis model for the given dataset and task.

By rigorously evaluating the performance of various sentiment analysis models using established evaluation metrics, the project ensures methodological robustness and facilitates informed decision-making regarding the selection of the most suitable model for sentiment analysis tasks in social media contexts.

X. RESULTS

1) ARVIND kEJRIWAL The sentiment analysis of Hindi-speaking audiences' tweets about Arvind Kejriwal shows a significant proportion of Neutral sentiment (43.9 per). This indicates a diverse range of opinions without a clear Positive or Negative sentiment dominance. Recommendations include conducting further sentiment trend analysis, monitoring sentiment trends, and adapting communication strategies to effectively manage public perception.
2) RAHUL GANDHI The sentiment analysis of posts about Rahul Gandhi among Hindi-speaking audiences reveals a significant dominance of Negative sentiment (40.9 percent). This indicates potential challenges in public perception, highlighting the importance of understanding Positive and Neutral sentiment proportions for a comprehensive assessment. Recommendations include conducting deeper sentiment analysis and developing targeted engagement strategies to address Negative sentiment while fostering Positive and Neutral sentiments.
3) NARENDRA MODI The sentiment analysis of posts about Narendra Modi among Hindi-speaking audiences shows a substantial prevalence of Negative sentiment (42.4 percent). This indicates potential challenges in public perception, emphasizing the need for detailed analysis of Positive and Neutral sentiments to devise effective engagement strategies.

Similarly there are results evaluated for all leader in Gujarati, Bengali, Telugu and Punjabi. Each language speaking data has different results for the respective leaders specified in the particular graphs for each.

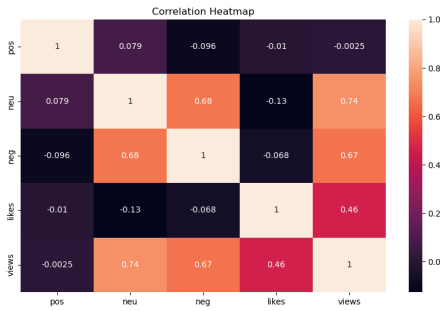

Fig. 1. Correlation Heatmap for Arvind Kejriwal

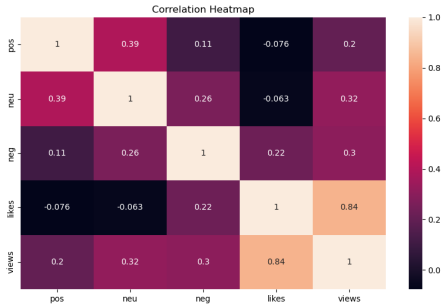

Fig. 2. Correlation Heatmap for Rahul Gandhi

XI. CONCLUSIONS

The project successfully aimed to provide a comprehensive sentiment analysis framework for ShareChat posts, incorporating both textual and visual data. By leveraging a series of modules and tools, including ShareChat scraper, Tesseract for image-to-text extraction, DeepTranslator for multilingual translation, and NLTK's SentimentAnalyzer trained on Vader Lexicon, the project facilitated the extraction and analysis of sentiments from a diverse range of posts.

Through manual labeling and comparison with various sentiment analysis models, including Transformer-based models, Flair, TextBlob, and VADER, the project evaluated the performance of each model in capturing sentiments accurately. The evaluation metrics employed, such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC value, provided a comprehensive assessment of model efficacy.

Furthermore, the incorporation of view and like counts to compute a mean normalized sentiment score offered a nuanced understanding of sentiment trends across different leaders. The user interface developed allowed for intuitive visual analysis, enabling users to explore sentiment scores for each leader interactively.

So, the project not only addressed the challenges of sentiment analysis on ShareChat but also provided a robust framework for sentiment analysis on social media platforms in general.

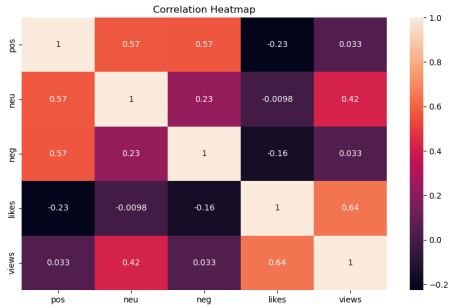

Fig. 3. Correlation Heatmap for Narendra Modi

REFERENCES

[1] A. Pathak, S. Kumar, P. P. Roy, and B.-G. Kim, "Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models," Electronics, vol. 10, no. 21, Art. no. 21, Jan. 2021, doi: 10.3390/electronics10212641.

[2] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," IEEE Access, vol. 11, pp. 15996–16020, Jan. 2023, doi: 10.1109/ACCESS.2022.3224136.

[3] K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognit Comput, vol. 8, pp. 757–771, 2016, doi: 10.1007/s12559-016-9415-7.

[4] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, "Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data," Neural Comput Applic, vol. 35, no. 29, pp. 21415–21431, Oct. 2023, doi: 10.1007/s00521-023-08629-3.