Journal of Marketing & Social Research ISSN (Online): 3008-0711
Volume: 03 | Issue 02 | 2026
Journal homepage: https://jmsr-online.com/
Research Article

Sentiment Analysis of Regional Political Discourse in Indian Languages: Evidence from Bihar Assembly Election 2025

***Swati Arya[1], Dr. Ankush Mittal[2] and Dr. Amit Agarwal[3]**
[1]Research Scholar Computer Science and Engineering COER University, Roorkee Uttarakhand
[2]Professor Computer Science and Engineering COER University, Roorkee Uttarakhand
[3]Scientist Research and Development Wells Fargo Banglore, Karnataka, India.

Received: 12/01/2026; Revision: 17/01/2026; Accepted: 11/02/2026; Published: 26/02/2026

Corresponding author: Swati Arya *(swati.arya.aiml@coeruniversity.ac.in)

Abstract: Sentiment analysis is the task of identifying the affective orientation of textual content and classifying it into predefined sentiment categories such as positive, negative, or neutral. With the rapid growth of social media usage in India, large volumes of politically relevant and code-mixed Hindi–English text are generated, particularly during regional elections. The Bihar Assembly Election 2025 witnessed extensive online discussions containing informal, transliterated, and context dependent political expressions. Accurately interpreting such code-mixed discourse is essential for understanding digital public opinion. This work presents an NLP-based framework for analysing approximately 167,000 Reddit comments related to the Bihar Assembly Election 2025. The proposed framework performs sentiment polarity classification along with fine-grained emotion extraction, including trust, anticipation, anger, and fear. Experimental results reveal distinct sentiment distributions across political parties and increasing sentiment volatility during the campaign phase. The study demonstrates the effectiveness of language-aware NLP techniques for modelling regional political discourse in Indian code-mixed social media settings.

Keywords: Digital Political Behaviour, Public Sentiment, Natural Language Processing, Bihar Assembly Election 2025, Code-Mixed Text, Social Media Analytics.

INTRODUCTION

Indian political discourse on social media is characterized by linguistic diversity, informal grammar, political jargon, and frequent code-mixing between English and regional languages. These characteristics present unique challenges for Natural Language Processing (NLP), particularly in sentiment analysis tasks. Regional elections such as the Bihar Assembly Election 2025, contested primarily by major alliances including the National Democratic Alliance (NDA) led by Bharatiya Janata Party (BJP) and Janata Dal (United) [JD(U)], and the opposition alliance led by Rashtriya Janata Dal (RJD) and the Indian National Congress (INC), offer a valuable case study for applying NLP techniques to politically rich, user-generated Indian-language text. Online discussions frequently reference party alliances, leadership figures, campaign slogans, and governance issues using informal and code-mixed expressions. Unlike monolingual text, such discourse combines English and regional languages often written in Roman script within the same sentence creating challenges such as transliteration inconsistency, lexical ambiguity, and frequent language switching that complicate contextual sentiment interpretation.

With the rapid expansion of social media usage in India, online platforms have become major spaces for political discussion, particularly during regional elections. A significant portion of Indian users are fluent in both English and Hindi, resulting in frequent use of code-mixed Hinglish text, where Hindi expressions are written in Roman script alongside English words. In natural language processing (NLP), sentiment analysis of such code-mixed political discourse presents substantial challenges due to transliteration inconsistencies, informal grammar, and contextual ambiguity.

The Bihar Assembly Election 2025 generated extensive user-driven political discussions across digital platforms. This study aims to analyze sentiment and emotional dynamics in approximately 167,000 Reddit comments related to the election. A language-aware preprocessing pipeline was implemented to handle noisy, code-mixed Hinglish text, followed by sentiment polarity classification and emotion extraction. The proposed framework evaluates hybrid modeling approaches to capture nuanced political discourse patterns across alliances and campaign phases.

The paper presents related work, followed by dataset, preprocessing, and the proposed model. It then discusses the experimental results and analysis, and concludes with key findings and insights.

LITERATURE SURVEY

Code-mixed sentiment analysis gained structured attention through SemEval-2020 Task 9 [1], where ensemble neural architectures demonstrated competitive performance on

Name: Swati Arya
Email: swati.arya.aiml@coeruniversity.ac.in

Hinglish text [2]. Subsequent work on Hindi–English offensive and sentiment detection [3] and surveys of codeswitching datasets [4] underline the structural complexity of multilingual inputs. Theoretical modeling of synthetic code-mixed data [5] and corpus development efforts such as L3Cube-HingCorpus [6] further strengthen this research direction.

Transformer-based architectures, beginning with BERT [7], have significantly improved contextual modelling. Indian language extensions such as IndicNLPSuite [8] and MuRIL [9] enhance multilingual representation learning in low-resource settings. In political NLP, sentiment extracted from social media has been linked to electoral outcomes [10], polarization dynamics [11], and Indian election forecasting [15].

Foundational work in sentiment analysis [12], [13] and neural opinion mining [14] provides theoretical grounding for polarity classification and affect modeling. The development of multilingual datasets such as DravidianCodeMix [16] further expands code-mixed sentiment research to other Indian languages.

Language identification in code-switched data [17], transliteration mining tasks in FIRE [18], and cross-lingual embedding models [19] provide essential preprocessing and representation strategies for mixed-language text. Large-scale unsupervised multilingual representation learning [20] offers scalable solutions for low-resource contexts. Word-level tagging [21] and neural language models for code-switched text [22] further support robust modelling of multilingual discourse.

Beyond linguistic modelling, computational analyses of social media discourse [23] and transformer-based hate speech detection [24] demonstrate the importance of contextual robustness in socially sensitive domains. Alternative pretraining strategies such as ELECTRA [25] highlight continued advancements in transformer-based language modelling.

Despite these developments, comprehensive benchmarking of hybrid classical and transformer architectures on largescale Hindi–English political discourse remains limited. This study addresses that gap through systematic evaluation in a regional Indian election context.

NLP FRAMEWORK FOR SENTIMENT ANALYSIS OF INDIAN POLITICAL DISCOURSE

Political discourse on Indian social media presents a distinct and challenging linguistic environment for NLP. Unlike well-structured formal text, election-related discussions are characterized by informal syntax, political jargon, abbreviations, rhetorical expressions, and frequent code-mixing between English and regional languages. During regional elections such as the Bihar Assembly Election, online users often combine English with Hindi or region-specific political terminology, creating complex linguistic patterns that are difficult to process using standard NLP pipelines. This study proposes an NLP framework specifically designed to analyse sentiment in such politically contextualized Indian texts.

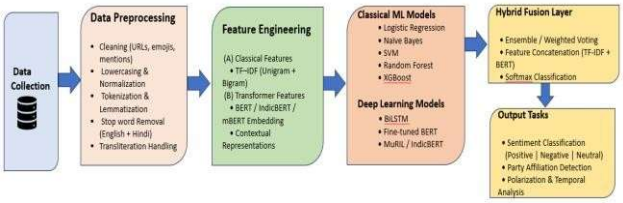

Figure 1: Hybrid ML-Transformer pipeline for codemixed political sentiment analysis.

The framework shown in **Figure 1** integrates TF–IDFbased classical machine learning models with contextual transformer embeddings (BERT/IndicBERT), followed by ensemble fusion for robust sentiment and party affiliation classification in multilingual political discourse.

Data Preparation and Linguistic Preprocessing

The raw textual data collected from the Reddit platform to extract political discussions, as shown in Figure 2, underwent a multi-stage preprocessing pipeline adapted to Indian political discourse. Initial cleaning steps removed URLs, emojis, special characters, and redundant whitespace while preserving politically meaningful tokens such as party names, alliance acronyms, and leadership references. Unlike generic text normalization approaches, domain-specific stop words related to elections like “vote”, “poll”, “seat” were selectively retained, as their removal was found to distort contextual sentiment interpretation.

Tokenization has been applied to segment text into meaningful lexical units, followed by lemmatization to reduce inflected forms to their base representations. Special attention has given to handling code-mixed expressions and transliterated words commonly used in Indian political discussions. For instance, Hindi terms written in Roman script were preserved to maintain semantic consistency rather than forcefully translated or removed. This approach ensured that culturally embedded political expressions remained intact throughout the analysis.

Sentiment Representation and Feature Extraction

To capture sentiment in informal and low-resource linguistic contexts, a lexicon-based sentiment analysis approach has been employed. Lexicon-based methods are particularly suitable for Indian political text due to their robustness against noisy language, spelling variations, and the limited availability of annotated regional datasets. Each text instance has assigned a sentiment polarity score representing positive, negative, or neutral orientation.

Beyond polarity classification, we extract fine-grained emotions such as trust, anticipation, anger, and fear to capture deeper affective signals in political discourse. This emotion-aware modeling enables a more nuanced

understanding of voter expectations, uncertainty, and dissatisfaction beyond explicit sentiment labels.

Handling Informality and Political Jargon

Indian political discourse frequently employs abbreviations, slogans, and context-dependent terminology such as party alliances, leader nicknames, and campaign phrases. To address this, a custom political vocabulary has been integrated into the NLP pipeline. This vocabulary included region-specific party names, alliance identifiers, and frequently used political terms relevant to the Bihar election. Retaining such expressions improved sentiment consistency and reduced misclassification caused by out-of-context lexical interpretation.

Moreover, the framework accounted for rhetorical structures commonly observed in political discussions, such as sarcasm, exaggeration, and evaluative comparisons. While detecting sarcasm remains a known challenge in sentiment analysis, maintaining surrounding contextual cues helped reduce polarity inversion errors in strongly opinionated texts.

Analytical Design and Sentiment Aggregation

Sentiment analysis has been conducted at both the individual text level as well as the aggregated group level. Individual posts and comments were first classified independently, after which sentiment distributions were aggregated based on political affiliation and election phase. This aggregation enabled comparative analysis of linguistic sentiment patterns across ruling alliances, opposition groups, and non-partisan discourse.

Temporal aggregation has also been employed to observe how sentiment and emotional intensity evolved throughout different stages of the election cycle. This design allowed the framework to capture shifts in public mood driven by campaign developments, political events, and perceived uncertainty.

METHODOLOGY USED

This study proposes a hybrid NLP-based analytical framework for modeling sentiment dynamics in Hindi–English code-mixed political discourse during the Bihar Assembly Election 2025. The methodology consists of data acquisition, language-aware preprocessing, sentiment and emotion modeling, temporal aggregation, and predictive analysis.

Data Acquisition and Corpus Construction

Political discourse data were collected from publicly accessible Reddit discussions using the official Reddit API. Data extraction covered the pre-election and campaign phases of the Bihar Assembly Election 2025. Posts and comments were filtered using a curated keyword lexicon comprising political party names, alliance identifiers, leader references, and election-specific terms.

Metadata including timestamps, engagement indicators (upvotes, reply counts), and thread identifiers were preserved to enable temporal and interaction-based analysis. After filtering and cleaning, the final corpus consisted of approximately 167,000 comments, containing English, Hindi (Romanized), and code-mixed Hinglish expressions**.**

Linguistic Preprocessing and Code-Mixed Handling

Given the informal and multilingual nature of social media discourse, a multi-stage preprocessing pipeline was implemented:

1. Noise Removal: URLs, HTML tags, emojis, excessive punctuation, and non-linguistic artifacts were removed.
2. Tokenization: Word-level tokenization was applied using a language-aware tokenizer.
3. Normalization: Elongated words and orthographic variations were standardized.
4. Transliteration Handling: Romanized Hindi tokens were preserved rather than translated to maintain contextual fidelity. Common transliteration variants were normalized to reduce lexical sparsity.
5. Lemmatization: English tokens were lemmatized, while Hindi tokens were retained in normalized surface form.

A domain-specific political lexicon containing party names, alliance codes, slogans, and leadership identifiers was integrated to reduce out-of-vocabulary errors and improve contextual interpretation.

Sentiment and Emotion Modeling

Sentiment polarity was computed using a lexicon-based scoring framework adapted for noisy and low-resource text environments. Each comment was assigned a polarity score $s_i \in \{-1,0,1\}$ representing negative, neutral, or positive orientation.

To capture deeper affective dimensions, emotion-level features were extracted using an affective lexicon framework. Emotional categories included trust, anticipation, anger, and fear. For each comment, emotion intensity scores were computed as normalized frequency-weighted sums of emotion-bearing tokens.

This dual-layer representation (polarity + emotion vectors) enabled richer modelling of political discourse beyond surface-level sentiment classification.

Temporal Aggregation and Sentiment Volatility

To analyse sentiment evolution over time, daily average sentiment was computed as:

$$S_t = \frac{1}{n_t}\sum_{i=1}^{n_t} s_i$$

where **"n_t"** denotes the number of comments on day **"t".** To reduce short-term noise and capture sustained trends,

rolling averages were computed over 7-day and 30-day windows:

$$S_t^{(k)} = \frac{1}{k} \sum_{j=t-k}^{t} S_j$$

Sentiment volatility was quantified using rolling standard deviation measures to assess fluctuations in public mood during different election phases.

Party-Level and Alliance-Level Aggregation

Comments were tagged based on the presence of political entity mentions. Sentiment distributions were computed for each party and alliance group. Mean sentiment and emotion scores were aggregated to enable comparative analysis across ruling and opposition alignments. This aggregation strategy allowed examination of asymmetry in digital narratives and emotional intensity across political blocs.

Predictive Modelling and Feature Importance

To examine relationships between sentiment features and engagement metrics, multiple supervised learning models were employed, including Linear-Regression, Ridge Regression, Random Forest, and Gradient Boosting.

Feature sets included:

- Polarity scores
- Emotion intensity vectors
- Rolling sentiment averages
- Party mention frequencies
- Engagement indicators
- Text length features

Tree-based ensemble models were used to capture nonlinear interactions. Feature importance was computed using impurity-based measures to identify the most influential predictors.

Engagement–Sentiment Relationship Analysis

The relationship between sentiment polarity and user engagement was examined using correlation and regression analysis. Engagement scores were plotted against sentiment values to identify patterns and trends. The objective was to assess whether extreme sentiments, either highly positive or negative, lead to greater interaction compared to moderate or neutral tones. The findings offer insights into how emotional intensity influences user behavior, participation, and overall responsiveness across different forms of content and communication.

RESULTS AND DISCUSSIONS

The empirical analysis reveals that sentiment volatility increases significantly as the election period approaches (**Figure 2**), indicating heightened polarization and intensified political engagement in online discourse. Party-level aggregation (**Figure 3**) demonstrates asymmetric sentiment patterns, suggesting that digital narratives vary substantially across political groups. From a modeling perspective, ensemble-based and transformer architectures outperform linear methods (**Figure 4**), confirming that political discourse in code-mixed environments exhibits non-linear and context-dependent relationships. Rolling sentiment averages emerge as the most influential predictors (**Figure 5**), highlighting the importance of temporal smoothing in noisy social media data. Importantly, engagement patterns (**Figure 7**) show that highly interactive posts are not necessarily driven by extreme polarity but often emerge from moderately negative or debate-oriented content. This suggests that contextual nuance and controversy play a stronger role in digital political interaction than purely positive messaging.

The findings demonstrate that integrating temporal aggregation, language-aware preprocessing, and hybrid modeling provides a robust framework for analyzing Hindi-English code-mixed political discourse. These results reinforce the need for NLP systems that account for linguistic diversity, contextual polarity shifts, and informal political expressions in Indian social media environments.

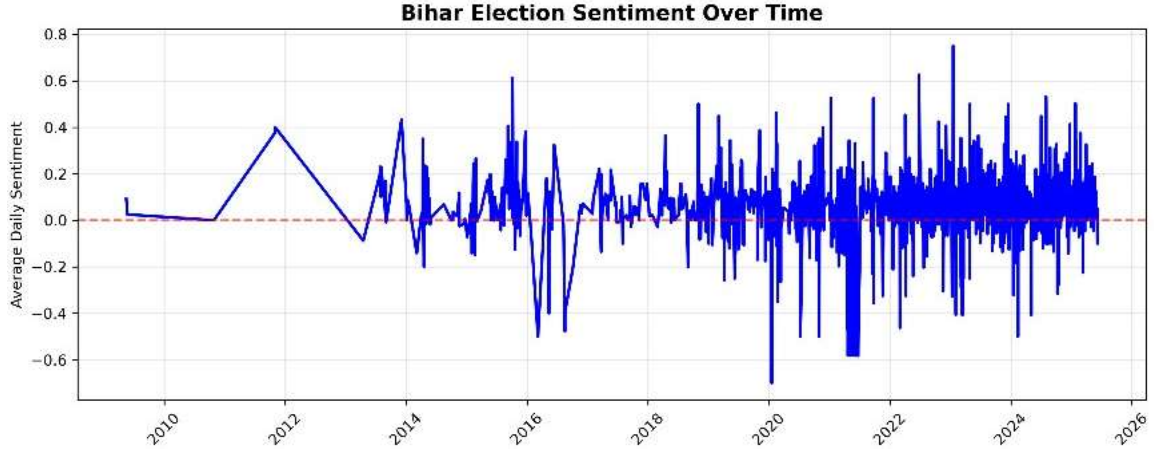

Figure 2: Temporal Evolution of Sentiment During Bihar Assembly Election 2025

Figure 2 presents the average daily sentiment score over time. The increasing volatility near the election period indicates rising political polarization and intensified digital engagement.

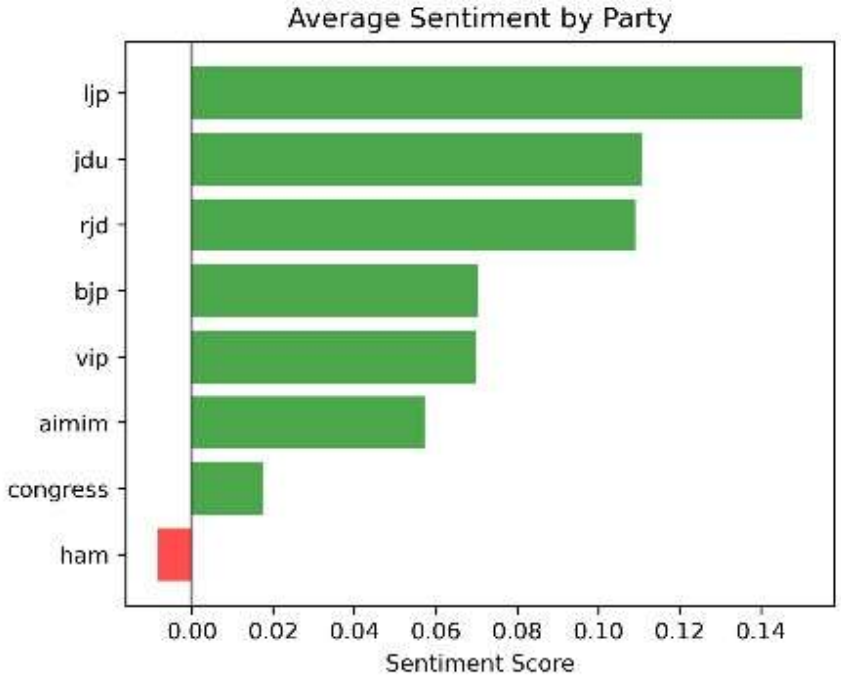

Figure 3: Average Sentiment Distribution Across Political Parties.

In **Figure 3**, horizontal bars in the chart illustrates mean-sentiment polarity by party affiliation. Distinct variations reflect asymmetric public perception and digital narrative differences among alliances.

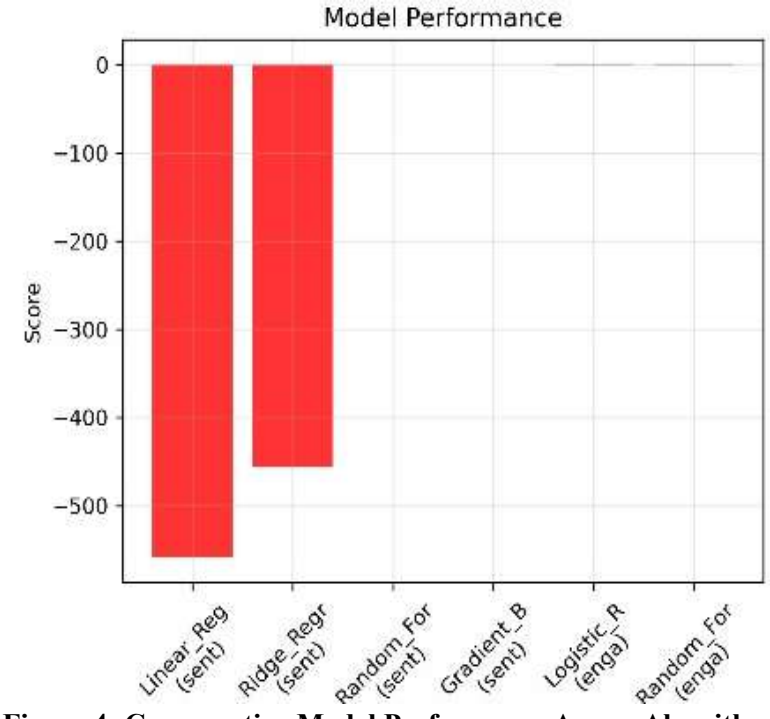

Figure 4: Comparative Model Performance Across Algorithms

Figure 4 compares the predictive performance of linear, ensemble, and transformer-based models. Non-linear ensemble approaches demonstrate superior predictive capability over linear regression methods.

Figure 3 shows sentiment differences across parties, indicating perception asymmetry, while **Figure 4** highlights superior predictive performance of ensemble models.

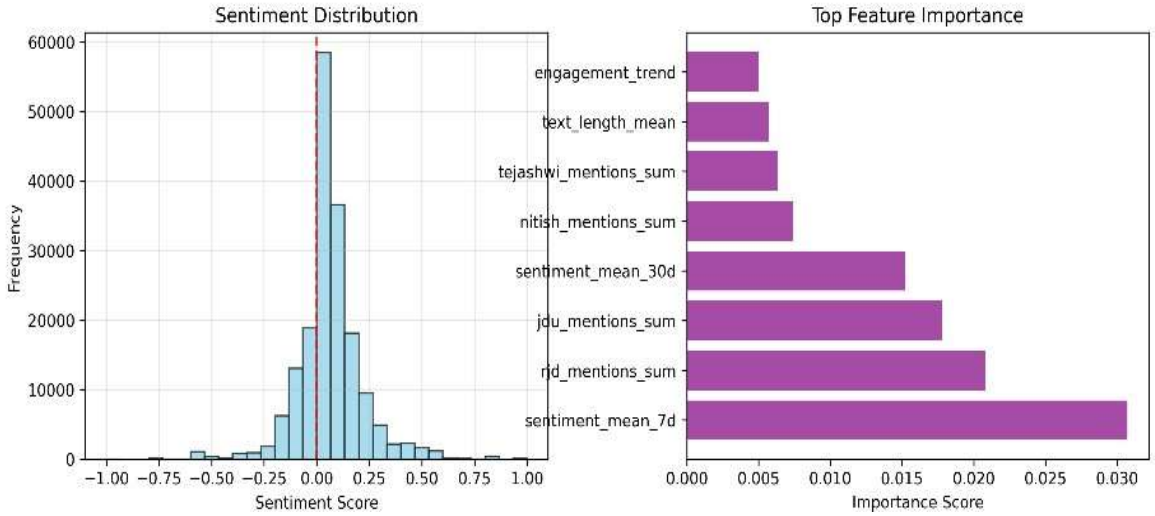

Figure 5: Top Feature Importance in Predictive Modeling.

Figure 5 shows feature importance analysis that highlights rolling sentiment averages and party mention frequencies as the most influential predictors, underscoring the value of temporal aggregation.

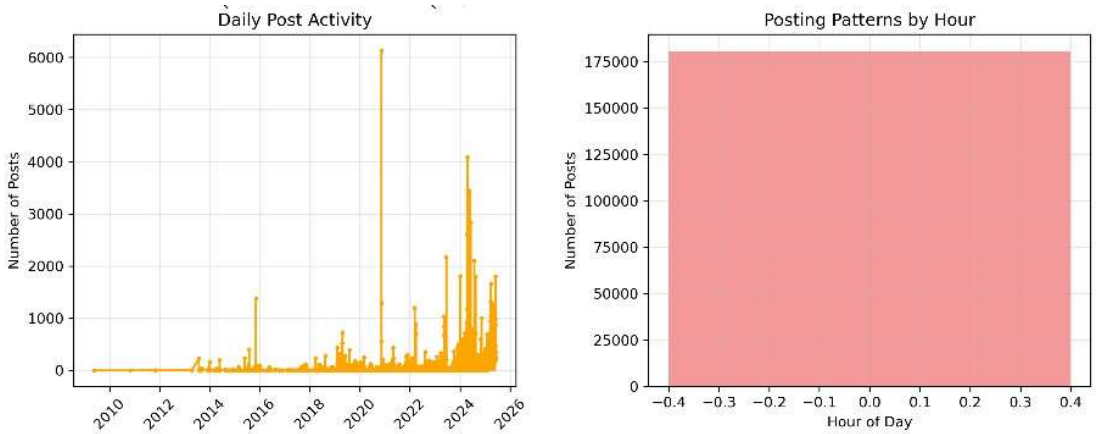

Figure 6: Sentiment Distribution and Daily Posting Activity.

The histogram in **Figure 6** illustrates overall sentiment distribution, showing dominance of neutral discourse with moderate polarity spread. The daily post activity plot reveals spikes corresponding to politically significant events and campaign phases.

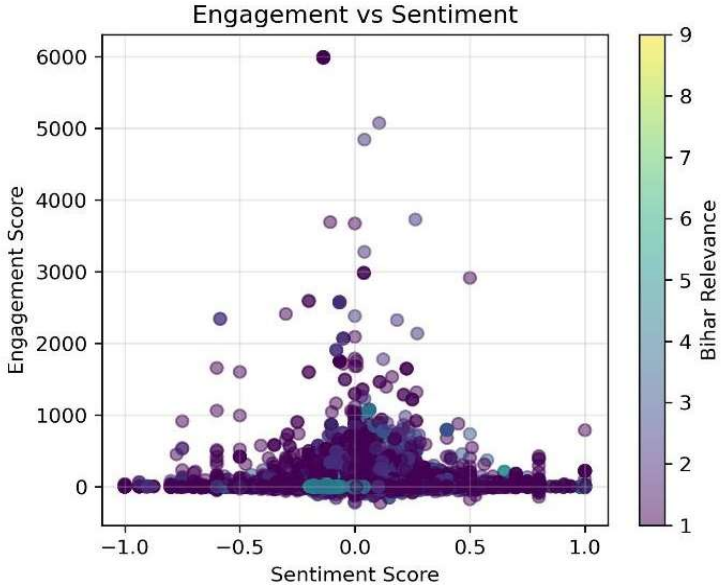

Figure 7: Engagement vs Sentiment Relationship.

The scatter plot in **Figure 7** depicts engagement scores against sentiment polarity. High engagement is concentrated around moderately negative or debate-oriented posts rather than extreme sentiment values.

The analysis reveals a mixed sentiment environment during the Bihar Assembly Election. Neutral sentiment dominates a large portion of the discourse, indicating analytical and informational discussions. Positive sentiment reflects optimism and support, while negative sentiment captures dissatisfaction and criticism.

DISCUSSIONS

The findings demonstrate that digital political sentiment provides meaningful insights into voter behaviour during regional elections. Rather than a binary positive-negative divide, the discourse reflects nuanced emotional states shaped by political alignment and perceived uncertainty. The prominence of anticipation highlights elections as emotionally charged events where expectations play a central role. Differences in emotional expression across alliances suggest varying degrees of confidence, polarization, and issue-based engagement. Importantly, neutral users exhibit lower emotional intensity, indicating a deliberative segment of the electorate that engages analytically rather than affectively.

From a behavioural perspective, these patterns underscore the role of digital platforms as spaces for collective sensemaking during democratic processes. Social media sentiment thus functions not merely as opinion expression but as an indicator of public trust and democratic engagement.

CONCLUSION

This work demonstrates that digitally expressed sentiment during the Bihar Assembly Election offers valuable insights into voter behaviour, emotional dynamics, and political engagement. By shifting focus from prediction to behavioural understanding, the study highlights the role of social media as a lens into democratic sentiment in regional contexts. The findings reinforce the importance of computational approaches for analysing public mood and contribute to interdisciplinary research at the intersection of political behaviour, data analytics, and governance.

REFERENCES

1. Patwa, Parth, et al. “SemEval-2020 Task 9: Sentiment Analysis of Code-Mixed Tweets.” *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval)*, 2020, pp. 774–790. https://doi.org/10.18653/v1/2020.semeval-1.99.
2. Baroi, S. J., et al. “NITS-Hinglish-SentiMix at SemEval-2020 Task 9: Sentiment Analysis for CodeMixed Social Media Text Using an Ensemble Model.” *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval)*, 2020, pp. 1298–1303. https://doi.org/10.18653/v1/2020.semeval-1.188.
3. Mathur, Puneet, et al. “Detecting Offensive Tweets in Hindi-English Code-Switched Language.” *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, 2018, pp. 18–26. https://doi.org/10.18653/v1/W18-3504.
4. Jose, Navya, et al. “A Survey of Current Datasets for Code-Switching Research.” *Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 136–141. https://doi.org/10.1109/ICACCS48705.2020.9074261
5. Pratapa, Adithya, et al. “Language Modeling for CodeMixing: The Role of Linguistic Theory-Based Synthetic Data.” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. https://doi.org/10.18653/v1/P18-1147.
6. Bansal, Aditi, et al. “L3Cube-HingCorpus: A Hinglish Code-Mixed Corpus for Sentiment Analysis.” *Proceedings of the 6th Workshop on Noisy User-Generated Text (WILDRE)*, 2022. https://doi.org/10.18653/v1/2022.wildre-1.2.
7. Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of NAACL-HLT*, 2019. https://doi.org/10.18653/v1/N19-1423.
8. Kakwani, Divyanshu, et al. “IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-Trained Multilingual Language Models for Indian Languages.” *arXiv*, 2020, arXiv:2009.10297.
9. Khanuja, Simran, et al. “MuRIL: Multilingual Representations for Indian Languages.” *Proceedings of EMNLP*, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.384.
10. Tumasjan, Andranik, et al. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.” *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2010. https://doi.org/10.1609/icwsm.v4i1.14009.
11. Xiong, Xinyu, et al. “Political Polarization on Social Media: A Computational Analysis.” *Computational Social Networks*, vol. 8, no. 1, 2021. https://doi.org/10.1186/s40649-021-00085-8.
12. Pang, Bo, and Lillian Lee. “Opinion Mining and Sentiment Analysis.” *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, 2008, pp. 1–135. https://doi.org/10.1561/1500000011.
13. Liu, Bing. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
14. Cambria, Erik, et al. “New Avenues in Opinion Mining and Sentiment Analysis.” *IEEE Intelligent Systems*, vol. 28, no. 2, 2013, pp. 15–21. https://doi.org/10.1109/MIS.2013.30.

15. Sharma, Ankit, et al. "Social Media Sentiment and Electoral Outcomes: Evidence from Indian Elections." *Decision Support Systems*, vol. 176, 2024. https://doi.org/10.1016/j.dss.2023.114049.
16. Chakravarthi, Bharathi Raja, et al. "DravidianCodeMix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text." *Language Resources and Evaluation*, vol. 56, 2022, pp. 765–806. https://doi.org/10.1007/s10579-022-09583-7.
17. Solorio, Thamar, et al. "Overview for the First Shared Task on Language Identification in Code-Switched Data." *Proceedings of the EMNLP Workshop on Computational Approaches to Code Switching*, 2014. https://doi.org/10.3115/v1/W14-3907.
18. Choudhury, Monojit, et al. "The FIRE 2013 Shared Task on Transliteration Mining." *Proceedings of FIRE*, 2013.
19. Ruder, Sebastian, et al. "A Survey of Cross-Lingual Word Embedding Models." *Journal of Artificial Intelligence Research*, vol. 65, 2019, pp. 569–631. https://doi.org/10.1613/jair.1.11640.
20. Conneau, Alexis, et al. "Unsupervised Cross-Lingual Representation Learning at Scale." *Proceedings of ACL*, 2020. https://doi.org/10.18653/v1/2020.aclmain.747.
21. Reddy, Y. P. S., et al. "Word-Level Language Identification for Hindi-English Code-Mixed Social Media Text." *Proceedings of the EMNLP Workshop on Code Switching*, 2016.
22. Winata, Genta Indra, et al. "Code-Switched Language Models Using Neural Architectures." *Proceedings of NAACL-HLT*, 2019. https://doi.org/10.18653/v1/N191030.
23. Rudra, Koustav, et al. "Extracting Situational Awareness from Twitter during Disaster Events." *Proceedings of the WWW Companion*, 2015. https://doi.org/10.1145/2740908.2741714.
24. Mozafari, Marzieh, et al. "Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model." *PLoS ONE*, vol. 15, no. 8, 2020. https://doi.org/10.1371/journal.pone.0237861.
25. Clark, Kevin, et al. "ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.