When Neutral Turns Negative: Cross-Domain Failure Modes in Hinglish Political Sentiment Analysis

Chennuru Rahul[1], Kolawole Adebayo[1,2]

Rahul.chennuru.2026@mumail.ie, Kolawole.adebayo@mu.ie
ADAPT Centre, Computer Science Department, Maynooth University, Ireland[1]
Maynooth International Engineering College, Maynooth University, Maynooth, Ireland[2]

Abstract

Sentiment analysis models are increasingly deployed to analyze political discourse, yet strong in-domain performance does not guarantee robustness under domain shift. We study cross-domain generalization in Hinglish (Hindi–English code-mixed) sentiment analysis by evaluating a fine-tuned XLM-RoBERTa classifier, trained on 29,000 general-domain Hinglish sentences, on a curated benchmark of politically oriented Hinglish text. While the model achieves 92.02% accuracy in-domain, performance drops to 71.83% under political domain shift.

Error analysis reveals a pronounced directional bias with 48.9% of neutral political statements misclassified as negative, indicating a systematic neutrality-to-negative shift. In addition, 87.5% of incorrect predictions are assigned confidence scores above 95%, pointing to severe miscalibration under distribution shift. We further compare these results against an instruction-tuned large language model (Llama 3.3), which achieves 90.85% zero-shot accuracy and 94.37% accuracy with contextual prompting, while substantially reducing neutrality bias. Our findings indicate the need for domain-aware evaluation, calibration diagnostics, and explicit reporting of failure modes when deploying sentiment models in politically sensitive settings.

Keywords: Hinglish, Political Sentiment Analysis, Domain Shift, Neutrality Bias, Large Language Models

1. Introduction

Sentiment analysis for multilingual and code-mixed text has gained significant attention due to the prevalence of mixed-language communication on social media platforms. In particular, Hinglish, a code-mixed combination of Hindi and English written in Latin script, poses unique challenges for automatic sentiment classification. These challenges arise from non-standard orthography and transliteration variability, which increase lexical sparsity, as well as lexical borrowing and syntactic mixing, which complicate the identification and composition of sentiment cues. Transformer models such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) have provided strong performance on multilingual and code-mixed benchmarks, including SemEval shared tasks such as SemEval-2020 Task 9 (Patwa et al., 2020). More recent work continues to refine transformer-based approaches for code-mixed and low-resource settings through architectural adaptations and data-centric strategies (e.g., Hashmi et al., 2024; Ahmad et al., 2025).

However, most prior studies evaluate models under homogeneous topical conditions, where training and test data are drawn from similar distributions. As a result, high benchmark accuracy may obscure vulnerabilities under distribution shift. Robustness and calibration under domain shift have been shown to degrade in neural models, even when in-domain performance remains strong (Guo et al., 2017; Ovadia et al., 2019). Despite this, systematic investigations of cross-domain reliability in code-mixed political sentiment analysis remain limited.

Political discourse introduces additional complexity. Factual reporting, policy discussion, stance expression, and affective judgment are often intertwined, making sentiment boundaries less explicit than in domains such as product reviews or informal social media posts. Prior work on stance and political affect detection (e.g., Mohammad et al., 2016) suggests that political language frequently encodes evaluation indirectly, increasing the risk of misclassification, particularly between neutral and negative categories. In politically sensitive settings, such directional errors can distort downstream analyses of public opinion.

In parallel, large language models (LLMs) have demonstrated strong zero-shot and few-shot classification capabilities across domains. However, their robustness and calibration behaviour in politically oriented, code-mixed contexts remain underexplored, particularly in comparison to fine-tuned encoder-based models (Shen et al., 2025; Bashiri et. al., 2024).

In this paper, we investigate cross-domain robustness in Hinglish sentiment analysis by evaluating a fine-tuned XLM-RoBERTa classifier, trained on general-domain Hinglish data, on a curated benchmark of policy-oriented political discourse. We analyze both performance degradation and calibration behaviour under domain shift and compare results with an instruction-tuned LLM (Llama 3.3) in zero-shot and prompted settings.

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 219–227
16 May 2026, Palma, Mallorca (Spain).

We address the following research questions:

- **Cross-domain robustness:** How does the accuracy and error distribution of a fine-tuned Hinglish sentiment model change when evaluated on political discourse that differs from its training distribution?
- **Directional bias and calibration:** Does political domain shift induce systematic misclassification patterns, particularly neutrality-to-negativity bias, and how does model confidence behave under shift?
- **Model class comparison:** Do instruction-tuned large language models exhibit greater robustness and reduced bias under political domain shift compared to fine-tuned encoder-based models?

By systematically analyzing accuracy, error directionality, and confidence under distribution shift, this study contributes empirical evidence on the reliability of sentiment models in politically sensitive Hinglish contexts. Our findings highlight the importance of domain-aware evaluation protocols, calibration diagnostics, and transparent reporting of failure modes when sentiment systems are deployed for political discourse analysis.

Contributions. (1) We identify systematic directional bias and overconfidence under domain shift in Hinglish political sentiment analysis; (2) we provide a comparative evaluation of encoder-based and instruction-tuned LLM approaches under this setting; and (3) we show that instruction-guided prompting can mitigate bias and improve robustness.

2. Related Work

2.1 Code-Mixed and Hinglish Sentiment Analysis

Sentiment analysis for code-mixed text, particularly Hindi–English (Hinglish), has attracted increasing attention due to the prevalence of multilingual communication on social media platforms. Shared tasks such as SemEval-2020 Task 9 (Patwa et al., 2020) and multilingual benchmarks like LinCE (Aguilar et al., 2019) have provided standardized evaluation settings for sentiment and related tasks in code-mixed contexts. These datasets highlight challenges including transliteration variability, non-standard spelling, lexical borrowing, and syntactic mixing.

Transformer-based architectures, such as BERT (Devlin et al., 2019) and XLM-R (Conneau et. al., 2020), have shown strong performance on multilingual and code-mixed sentiment benchmarks. More recent studies extend this work through architectural adaptations, multilingual pretraining strategies, and low-resource augmentation methods (Hashmi et al., 2024; Ahmad et al., 2025). Other works examine model interpretability in code-mixed sentiment contexts, indicating broad interest in multilingual sentiment performance (Krasitskii et al., 2025; Xie, 2026). While these approaches improve in-domain performance, evaluation typically assumes similar topical and stylistic distributions between training and test data. Consequently, little attention has been paid to robustness under domain shift within code-mixed settings.

2.2 Domain Shift and Robustness in NLP

Distributional shift, where the test distribution differs from the training distribution in vocabulary, topic, or style, poses a well-documented challenge in NLP and machine learning. Empirical studies provide that neural models can suffer substantial performance degradation under such shifts, even when in-domain metrics remain high (Rabanser et al., 2019; Calderon et al., 2024). Robustness research in NLP has therefore emphasized evaluation on out-of-distribution (OOD) data and stress-testing across domains (Wang et al., 2024; Vajjala & Shimangaud, 2025).

Beyond accuracy degradation, calibration under shift is a critical concern. Neural classifiers are often poorly calibrated, exhibiting overconfident predictions even when incorrect. Prior works, such as Guo et al. (2017) demonstrate that modern neural networks can be substantially miscalibrated, and that calibration often worsens under distribution shift. However, systematic analysis of calibration behaviour in code-mixed political sentiment tasks remains limited. Our work extends robustness and calibration analysis to Hinglish sentiment classification under political domain shift.

2.3 LLM Generalization and Zero-Shot Performance

Instruction-tuned large language models (LLMs) have demonstrated strong zero-shot and few-shot generalization across a range of classification tasks. Models such as LLaMA series (Touvron et al., 2023) show competitive performance without task-specific fine-tuning, often leveraging prompting strategies to adapt to new domains. These capabilities suggest potential robustness advantages relative to fine-tuned encoder models, which may overfit to training distributions.

However, emerging evaluations reveal variability in cross-lingual and cross-domain transfer performance, particularly in low-resource or code-mixed contexts (Vajjala & Shimangaud, 2025; Shen et al., 2025). Systematic comparisons between encoder-based sentiment models and instruction-tuned LLMs under controlled political domain shift remain sparse, especially for

codemixed settings like Hinglish. Our study directly evaluates these model classes under identical cross-domain conditions.

2.4 Political Discourse, Sentiment & Stance

Political discourse presents unique challenges for sentiment classification. Policy discussion, factual narration, stance expression, and affective evaluation frequently co-occur within the same utterance, complicating categorical sentiment labelling. Prior work on political stance detection demonstrates that political language often encodes evaluation indirectly or implicitly (Mohammad et al., 2016).

While major works around NLP application to politics space have extensively examined stance and opinion mining, few studies integrate political semantics with code-mixed sentiment analysis. In particular, the interaction between political content and neutrality judgments in multilingual, code-mixed contexts has not been systematically examined. This gap motivates our analysis of directional misclassification patterns and bias, under political domain shift in Hinglish sentiment models.

This study lies at the intersection of code-mixed sentiment analysis, domain shift robustness, calibration research, and political NLP. Unlike prior work that emphasizes in-domain benchmark performance, we explicitly evaluate cross-domain reliability, analyze directional error patterns, and compare encoder-based and instruction-tuned models under politically oriented distribution shift.

3. Data and Methodology

3.1 Datasets

We use two datasets to evaluate cross-domain robustness in Hinglish sentiment classification.

3.1.1 General-Domain Training Data

The general-domain Hinglish dataset consists of approximately 29,000 sentences collected from publicly available Hinglish sentiment datasets released on GitHub, primarily derived from Twitter and other informal social media sources. The data reflects conversational and entertainment-oriented content rather than political discourse.

The dataset is labeled into three sentiment classes: positive, neutral, and negative, with the following distribution: negative (~50%), positive (~30%), and neutral (~20%). This skew toward negative sentiment results in relatively fewer neutral instances, which may influence model behaviour, particularly under domain shift.

In addition, lexical correlations in the training data, where certain topics or expressions co-occur with negative sentiment, may lead the model to associate topical cues with polarity. This provides a plausible explanation for the observed neutrality-to-negative bias when applied to political text.

3.1.2 Political Hinglish Benchmark

To evaluate cross-domain robustness, we construct a curated political evaluation benchmark consisting of 142 Hinglish sentences spanning 11 policy-relevant topics, including CAA/NRC, demonetization, farm laws, elections, budget, media, inflation, and foreign policy. The dataset is balanced across sentiment classes (Positive: 48; Negative: 49; Neutral: 45).

All instances were annotated independently by two native Hindi speakers with high proficiency in English and prior experience with sentiment labelling. Annotators were provided with written guidelines defining the three sentiment categories, with particular emphasis on distinguishing neutral factual statements from implicitly evaluative political commentary.

Inter-annotator agreement was measured using Cohen's κ, yielding $\kappa = 0.95$, indicating near-perfect agreement. Disagreements were resolved through adjudication in consultation with a third native Hindi speaker with linguistic training. The final dataset reflects the adjudicated consensus labels.

The benchmark is intended as a controlled diagnostic evaluation set for measuring performance under political domain shift rather than as a large-scale training corpus. To promote transparency and reproducibility, we release the benchmark publicly for research use at https://anonymous.4open.science/r/Datasets-8967.

All texts are Hindi–English code-mixed (Hinglish) written in Latin script. Compared to the training corpus, the political benchmark differs in (i) topical distribution (policy-oriented vs. conversational), (ii) vocabulary (governance and institutional terminology), and (iii) discourse structure (a higher prevalence of factual and policy-descriptive statements). This divergence enables controlled evaluation under domain shift without retraining.

3.2 Encoder-Based Model Fine-Tuning

We fine-tuned XLM-RoBERTa (Conneau et al., 2020) a multilingual transformer encoder model, for three-class sentiment classification. A linear classification head is added to the [CLS] representation, and the model is trained using cross-entropy loss. To avoid task-specific overfitting through extensive hyperparameter

tuning, we adopt a standard configuration of the HuggingFace Trainer framework, with a batch-size of 32 and a learning rate of 2e-5. The model is trained for 3 epochs, and the checkpoint with the best validation macro F1-score is selected. Evaluation on the general-domain dataset is conducted using accuracy and macro-averaged F1-score. This in-domain performance serves as a baseline against which cross-domain degradation on the political benchmark is measured.

3.3 Zero-Shot Classification

To assess robustness without task-specific fine-tuning, we evaluate LLaMA 3.3 (Touvron et al., 2023) in a zero-shot setting. The model is prompted to classify each input sentence as positive, neutral, or negative without access to labeled examples.

Although zero-shot at inference time, the model benefits from large-scale pretraining and instruction tuning. This evaluation tests whether general-purpose instruction alignment confers robustness under domain shift relative to supervised fine-tuning.

3.4 Context-Guided Prompting

We further evaluate LLaMA 3.3 using a context-guided prompting strategy designed to reduce neutrality bias under political domain shift. The prompt explicitly defines sentiment categories and instructs the model not to associate political topics with negative sentiment, emphasizing that factual or policy-descriptive statements should be labeled as neutral.

No labeled examples are provided, and no gradient-based updates are performed. This setting isolates the effect of task-definition framing on robustness and neutrality bias. The full prompt is provided in Appendix A. Decoding is carried out using a greedy strategy, with the temperature parameter set to zero.

3.5 Evaluation Under Distribution Shift

Domain shift in this study refers to the distributional divergence between the general-domain Hinglish training corpus and the curated political evaluation benchmark. Differences include topical focus, lexical distribution, and discourse structure. All models were evaluated using accuracy, macro F1-score, class-wise F1-scores, along with qualitative error analysis.

To assess reliability under shift, we further analyze prediction confidence for the encoder model using SoftMax probabilities. Specifically, we measure the proportion of incorrect predictions assigned a high confidence (≥ 95%) to quantify overconfidence.

Calibration is evaluated on the political benchmark to reflect model behaviour under domain shift.

Model	Acc %	Macro F1 %	Negative F1 %	Neutral F1 %	Positive F1 %
XLM-R (General)	92.02	85.01	94.18	67.69	93.17
XLM-R (Political)	71.83	70.93%	70.87	59.70	82.22
Llama 3.3 (Zero-Shot)	90.85	90.66%	94.00	87.06	90.91
Llama 3.3 (+Context)	94.37	94.33	95.92	93.18	93.88

Table 1: Overall Performance Across Domains and Models with 142 Political Benchmark Probes

Table 1 presents the overall results from our evaluation, showing better performance for in-domain scenarios. Figure 1 pinpoints the performance insight across domains and models.

4. Results

We evaluate cross-domain robustness by comparing model performance on the general-domain Hinglish test set and the curated political benchmark containing 142 probes. Performance metrics include accuracy, macro F1-score, class-wise F1-scores, and confidence-based error analysis. Results reveal substantial performance degradation for the fine-tuned encoder model under political domain shift, particularly neutral statements, alongside evidence of overconfident misclassification. In contrast, the instruction-tuned LLM demonstrates stronger robustness and improved handling of neutral political discourse.

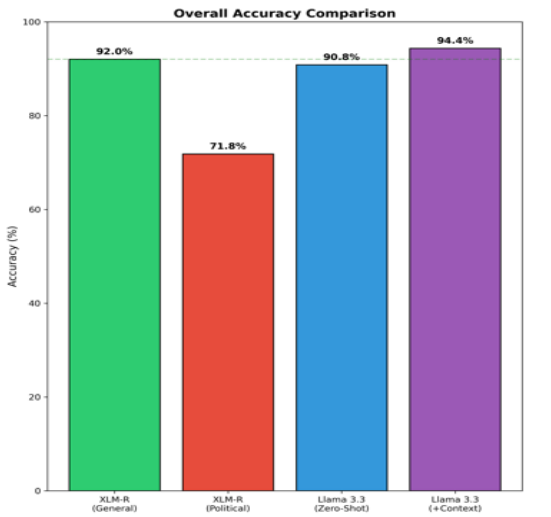

Figure 1: Overall accuracy comparison across domains and models. XLM-R shows substantial degradation under political domain shift, while LLaMA demonstrates stronger cross-domain robustness.

4.1 Cross-Domain Performance

Under political domain shift, the XLM-RoBERTa encoder's performance decreased substantially. For instance, accuracy dropped from 92.02% on general-domain Hinglish data to 71.83% on the political benchmark, and macro F1 declined from 85.01% to 70.93%. We provide the variance across the board in Table 2. Class-wise analysis reveals that neutral statements were most affected. For instance, neutral F1 decreased from 67.69% to 59.70%, becoming the weakest-performing category.

Metric	General Domain	Political Domain	Change
Accuracy	92.02%	71.83%	−20.19%
Macro F1	85.01%	70.93%	−14.08%
Neutral F1	67.69%	59.70%	−7.99%

Table 2: Performance degradation of XLM-RoBERTa under political domain shift.

Figure 2 compares per-class F1 scores across domains and models. The neutral category exhibits the largest relative degradation for XLM-R under political shift, whereas the LLaMA 3.3 maintains balanced performance across all classes. This disparity highlights the encoder model's sensitivity to neutrality distinctions in politically contextualized text.

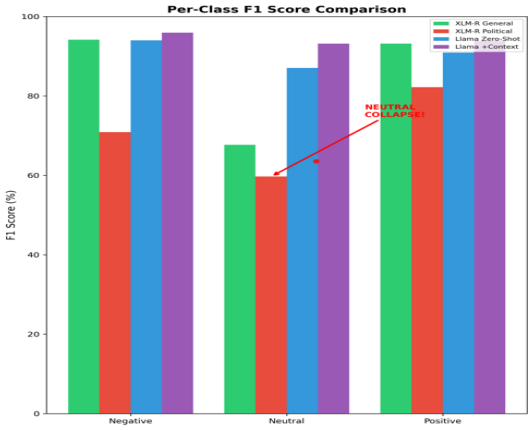

Figure 2: Per-class F1-score comparison across domains and models. Neutral performance degrades for XLM-R under political shift, while LLaMA maintains more balanced class performance

Moreover, error analysis indicates systematic directional misclassification. Figure 3 presents the confusion matrices for the general-domain and political benchmarks. Under domain shift, the confusion mass shifts noticeably toward the negative class, particularly for neutral instances. While class boundaries are well separated in the general domain, the political benchmark exhibits asymmetric confusion concentrated in neutral-to-negative predictions, reinforcing the presence of directional bias rather than uniform degradation.

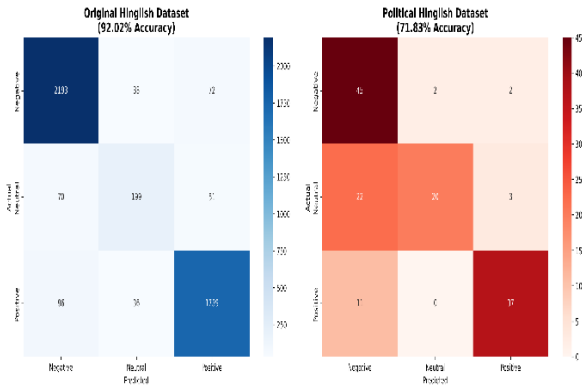

Figure 3: Confusion matrices for XLM-RoBERTa on the general-domain test set and the political benchmark. The political matrix shows increased neutral-to-negative misclassification under domain shift.

Error patterns suggest asymmetric degradation under domain shift. As shown in the confusion matrix in Figure 3, misclassifications are not uniformly distributed across classes but are

concentrated toward the negative class, particularly neutral instances. This indicates the presence of directional bias rather than random error.

Detailed analysis of confidence behavior and class-specific error patterns is provided in Section 4.2.

4.2 Error Analysis and Calibration

We analyze XLM-RoBERTa's prediction confidence on the political benchmark.

We observe a strong pattern of **overconfidence in incorrect predictions**. Out of 40 misclassified instances, 35 (87.5%) were assigned confidence ≥ 95%. Many of these errors are made with near-certainty, often approaching 99-100%.

This overconfidence is particularly evident in neutrality-related errors. Factual political statements such as:

- "*CAA ke under 6 religions ke refugees ko citizenship milegi…*"
- "*Aadhaar card ab 130 crore se zyada Indians ke paas registered hai…*"
- "*India ne UN mein abstain vote kiya…*"

are consistently misclassified as negative with confidence scores between 99–100%, despite being labelled as neutral.

This indicates a severe mismatch between predicted confidence and empirical correctness, suggesting that the model is poorly calibrated under political domain shift.

Moreover, additional analysis reveals a strong directional misclassification pattern. Out of 45 neutral political statements:

- 22 (48.9%) are misclassified as negative
- 3 (6.7%) are misclassified as positive
- 20 (44.4%) are correctly classified

This asymmetry indicates a systematic neutral-to-negative bias rather than uniform performance degradation.

To deepen our investigation, we further analyze cases where the instruction-tuned LLaMA 3.3 model succeeds while XLM-R fails. Out of 142 samples, 37 instances are correctly classified by LLaMA but misclassified by XLM-R. These cases are factual or policy-descriptive statements, where LLaMA assigns neutral or positive labels while XLM-R predicts negative with high confidence.

This suggests that instruction-tuned models are better able to distinguish between topical content and sentiment polarity, thereby reducing neutrality bias under domain shift.

Overall, these findings indicate that domain shift induces not only performance degradation but also systematic bias and severe overconfidence, raising concerns about the reliability of fine-tuned sentiment models in politically sensitive contexts.

Category	Count	Percentage
Total Incorrect Predictions	40	100%
Errors > 95% Confidence	35	87.5%
Errors ≤ 95% Confidence	5	12.5%

Table 3: Confidence distribution of incorrect XLM-RoBERTa predictions on the political benchmark

4.3 Large Language Model Performance

In contrast, as shown in Table 1, the instruction-tuned LLaMA 3.3 model demonstrates stronger robustness under political domain shift. In zero-shot mode, it achieves 90.85% accuracy and 87.06% neutral F1, outperforming XLM-R on neutral statements. With context-guided prompting where we explicitly instruct the model to treat factual political statements as neutral, the performance further improves 94.37% accuracy and 93.18% neutral F1.

Further topic-level evaluation shows that contextual guidance helps LLaMA handle previously challenging topics, including CAA/NRC and demonetization, where XLM-R struggles, as evidenced in Table 4. This suggests that instruction tuning and task-specific contextual framing can reduce neutrality bias.

The full prompt used for contextual evaluation is provided in Appendix A.

Topic	XLM-R (%)	Llama (%)	Difference
CAA/NRC	50.0%	100.0%	+50.0%
Demon	57.1	100.0	+42.9
Budget	71.4	100.0	+28.6
Farm Laws	64.3	85.7	+21.4
Media	55.6	88.9	+33.3
Inflation	66.7	100.0	+33.3
Election	100.0	100.0	0.0

Table 4: Topic-level accuracy of models on the political benchmark.

4.4 Interpretation and Implications

The results highlight several key insights:

- *High in-domain performance is not sufficient*: The observed degradation under political domain shift demonstrates that high in-domain performance does not ensure robustness when models are exposed to shifts in topic and discourse structure. Although the encoder model performs strongly on general Hinglish content, it appears sensitive to lexical and topical cues that differ in political contexts.
- *Neutral statements are particularly vulnerable:* The pronounced drop in neutral F1 score suggests that the encoder model may over-rely on surface-level lexical associations learned during training. Political terms related to protest, governance, or controversy, while not inherently negative, may co-occur with negative sentiment in the general-domain corpus, leading the model to conflate **topic salience with polarity** under shift. This results in systematic neutrality-to-negative misclassification rather than random confusion.
- *Calibration issues amplify risk*: The high rate of overconfident errors further indicates calibration challenges. Particularly, in the case of XLM-R, the model not only misclassifies neutral political discourse but does so with strong confidence, reducing the reliability of probability scores as indicators of correctness. In applied settings, such as public opinion tracking, policy discourse analysis, or media monitoring, such systematic polarization of neutral content could distort aggregate sentiment estimates.
- *Instruction-tuned LLMs show greater contextual sensitivity*: Instruction-tuned LLM demonstrates greater contextual sensitivity, particularly in distinguishing factual policy statements from evaluative language. This suggests that large-scale pretraining as is the case with LLaMA 3.3, combined with instruction alignment may mitigate neutrality bias and reliance on narrow lexical heuristics for prediction. This in turn improves robustness to other domain-specific lingusitic biases.
- Contextual prompting helps**:** Explicit task definitions improve classification of neutral political statements, highlighting the importance of domain-aware evaluation and careful prompt engineering.

That said, topic-level results should be interpreted cautiously due to the limited number of samples per topic. The performance improvement with contextual prompting suggests that clearer task definition, particularly around neutral political statements, plays a key role in mitigating domain-induced bias.

Overall, these findings emphasize the importance of domain-aware evaluation and calibration diagnostics prior to deploying sentiment models in politically sensitive contexts. Without such safeguards, neutral political discourse may be systematically polarized, potentially biasing downstream analytical conclusions, especially in social or policy research which may influence government decisions.

5. Conclusion

This paper investigated the robustness of a fine-tuned Hinglish sentiment model under political domain shift. While the encoder-based XLM-RoBERTa model achieved 92.02% accuracy on general-domain Hinglish data, performance dropped to 71.83% on a curated political benchmark. Error analysis revealed directional misclassification, with half of neutral political statements labeled as negative, and most errors made with high confidence, indicating calibration issues under distributional shift.

In contrast, an instruction-tuned large language model demonstrated greater robustness, achieving higher overall accuracy and more reliable handling of neutral political statements, particularly when provided with explicit contextual guidance. These findings imply that high in-domain benchmark performance does not necessarily guarantee reliability in politically sensitive contexts.

Overall, our results highlight the importance of domain-specific evaluation, transparent reporting of model limitations, and careful validation before deploying sentiment analysis systems for political discourse analysis.

6. Limitations

While the curated political benchmark provides a diagnostic evaluation of domain shift, its small size limits statistical generalization. Topic-level analyses are constrained by low per-topic sample sizes, and only one encoder-based model (XLM-RoBERTa) and decoder LLM (LLaMA 3.3) were evaluated. We do not include standard calibration metrics such as Expected Calibration Error (ECE) or reliability diagrams. Consequently, our findings indicate trends in cross-domain behaviour rather than absolute performance bounds.

The political benchmark consists of 142 instances and is intended for controlled diagnostic evaluation rather than statistical generalization.

While it enables analysis of cross-domain behavior and error patterns, the small sample size limits the robustness of fine-grained conclusions.

Topic-level analyses are exploratory and may be sensitive to small sample effects. Future work should validate these findings on larger and more diverse political Hinglish datasets.

Future work should explore larger and more diverse political Hinglish datasets, incorporate additional model baselines spanning different architectures and training strategies, and include formal robustness and calibration evaluations to better quantify model uncertainty under domain shift.

7. Ethics Statement

This study uses publicly available political Hinglish text and does **not** involve personal user information, demographic profiling, or attempts to influence political opinion. The focus is on evaluating model behavior under domain shift rather than on political analysis or persuasion.

Given the sensitivity of political discourse, we caution that automated sentiment systems **should not be deployed without domain-specific validation and human oversight**. Our results are provided to promote transparency, reproducibility, and accountable assessment of NLP models in politically sensitive contexts.

References

A Garg. (2026). Code-Mix Sentiment Analysis on Hinglish Tweets. arXiv preprint.

Advani, L., Lu, C., & Maharjan, S. (2020). Code-mixed sentiment analysis with feature engineering. *ArXiv preprint*.

Aguilar, G., Kar, S., Solorio, T., & Das, A. (2019). LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL 2019)*, pages 180–195.

Barnes, J., Klinger, R., & Schulte im Walde, S. (2018). Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains. *EMNLP 2018*.

Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowledge and Information Systems*, *66*(12), 7305-7361.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 36-41).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8440–8451.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pages 4171–4186.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K.Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330.

Hashmi, M., Singh, A., & Verma, P. (2024). Advances in code-mixed sentiment analysis: Transformer-based approaches and benchmarks. *Journal of AI Language Research*, 12(4), 45–60.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of ICLR 2019*.

Kalaivani, K.S. (2025). Political multiclass sentiment analysis of Tamil X political tweets. *DravidianLangTech Workshop Proceedings*.

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Talukdar, P. (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Khatavkar, V. (2025). Multilingual transformer contextual embedding model for political tweets analysis. Technical report.

Li, Y., Smith, A., & Zhao, B. (2024). Evaluating large language models under distribution shift. *ACL 2024 Findings*.

LREC Workshop on Language and Politics 2024.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of SemEval-2016*.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., Gambäck, B., Chakraborty, T., Solorio, T., & Das, A. (2020). SemEval-2020 Task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, pages 774–790.

Rabanser, S., Günnemann, S., & Lipton, Z.C. (2019). Failing loudly: Empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 1396–1408.

Sankar, P.S.S. (2024). Sentiment analysis of code-mixed languages: Methods and challenges. *International Journal of Innovative Engineering and Emerging Technology*.

Sharma, P., & Singh, R. (2024). Analyzing political discourse in multilingual social media.

Shen, L., Gupta, R., & Lee, C. (2025). Political sentiment and stance detection in social media: Challenges and trends. *Journal of Computational Social Science*, 8(1), 112–130.

Singh, P., & Lefever, E. (2020). Cross-lingual embeddings for sentiment analysis of Hinglish social media text. *ArXiv preprint*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint.

Vajjala, S., & Shimangaud, R. (2025). Benchmarking generalization of instruction-tuned models across languages. *NAACL 2025*.

Wang, Q., Müller, M., & Liu, J. (2024). On the robustness of instruction-tuned language models. *EMNLP 2024*.

Appendix A: Prompt Used for LLaMA 3.3

You are a sentiment classifier for Hinglish (Hindi–English code-mixed) text.

Classify the sentiment of the given sentence as one of:
Positive, Neutral, or Negative.

Definitions:

- Positive: expresses happiness, praise, support, or positive emotion
- Negative: expresses criticism, anger, sadness, or negative emotion
- Neutral: factual statements, policy descriptions, or sentences without clear sentiment

Guidelines:

- Do not infer sentiment from topic alone
- Political or policy-related statements can be Neutral if they are descriptive
- Use Neutral when no clear opinion is expressed
- Sarcasm with negative intent should be labeled as Negative

Sentence: input}

Answer: