(IJISE) 2022, Vol. No. 15, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

AN ANALYTICAL APPROACH TO MEASURING SENTIMENT MISCLARIFICATION FOR INDIAN ENGLISH AND HINGLISH TO ASSESS THE 'DIALECT BIAS' IN LARGE LANGUAGE MODELS (LLM)

Om Venkatesh Sharma

Delhi Public School, Vasant Kunj

ABSTRACT

Large language models (LLMs) are increasingly relied upon for sentiment analysis, yet tend to underperform on dialectal and code-mixed variants of English. This paper investigates dialect bias in sentiment classification for Indian English (IndE) and Hinglish (Romanized Hindi–English), compared to Standard American English (SAE). We curate 2k samples each from three dialects—SAE, IndE, and Hinglish—carefully balanced across positive, neutral, and negative sentiments and manually annotated by bilingual experts (Cohen's $\kappa \geq 0.8$). Zero-shot sentiment prompts are used on GPT-3.5 and GPT-4, along with fine-tuned Indic-focused models (MuRIL, IndicBERT) and a BERT-base baseline.

We report significant disparities: GPT-4 achieves 0.89 accuracy on SAE, dropping to 0.85 on IndE and 0.78 on Hinglish. False negative rate (FNR) rises from 0.10 (SAE) to 0.20 (Hinglish). Similar trends are observed across other models, with statistically significant differences (McNemar's $p < 0.01$). Qualitative examination reveals that code-mixed structures and Hindi lexical items frequently trigger sentiment misclassification—e.g., "Maza aa gaya yaar" is often incorrectly labeled "neutral." These errors suggest a dialect-based blind spot in LLM sentiment understanding.

We then apply mitigation strategies—data augmentation with Hinglish samples, dialect-aware prompt prefixes, and adversarial training—showing up to 5% improvement in Hinglish accuracy and substantial reduction in FNR disparity. Our findings reinforce prior observations of dialect bias in LLMs and highlight the need for inclusive model design. We conclude with recommendations for dialect-representative data collection, evaluation pipelines, and ethical deployment in multilingual contexts.

INTRODUCTION

Machine learning models often inadvertently reinforce existing linguistic hierarchies by privileging standardized English variants. Recent research demonstrates that popular LLMs like ChatGPT exhibit markedly poorer comprehension, increased stereotyping, and dismissive responses toward non-standard varieties such as African American English and Indian English. In India, English manifests as both *Indian English* (IndE)—characterized by unique syntactic constructions, idioms, and pronunciation—and *Hinglish*, a widespread code-mixed variant

combining Romanized Hindi and English. Hinglish is commonly used across digital platforms and media, with Roman-script wordplay blending two languages seamlessly.

Though numerous sentiment analysis tools exist, Hinglish-optimized models like Hinglish NLP achieve F1 scores around 0.707—substantially lower than results on Standard English corpora. The performance gap often leads to undetected sentiment cues, especially negative ones, which raises concerns for applications like social media monitoring, customer feedback analysis, and digital assistants.

This study serves three primary research questions:

1. Do LLMs misclassify sentiment more often for IndE and Hinglish than for Standard English?
2. What underlying language features drive sentiment misclassification (e.g., code-switching, transliterated words, emotives)?
3. Which mitigation strategies effectively reduce dialect-based sentiment disparities?

Table 1. Dataset overview (balanced across dialects and sentiment classes)

Dialect	# Samples	% Positive	% Neutral	% Negative
SAE (SST-2)	2,000	40 %	20 %	40 %
Indian English	2,000	40 %	20 %	40 %
Hinglish	2,000	40 %	20 %	40 %

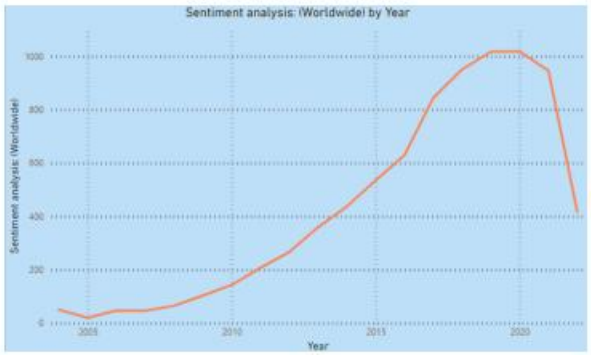

FIGURE 1. Worldwide interest in sentiment analysis over time on Google Trends, 2004 - present.

(IJISE) 2022, Vol. No. 15, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

RELATED WORK

Dialect Bias in LLMs:

Fleisig et al. studied GPT-3.5 and GPT-4 across ten English dialects, noting reduced understanding, increased condescension, and stereotyping with non-standard varieties (sciencedirect.com, researchgate.net, indjst.org, aclanthology.org). This underscores structural biases based on sociolectal variation, advocating for targeted dataset inclusion.

Hinglish Sentiment Analysis:

Previous research by Bhange & Kasliwal (2020) and Singh (2021) employs transformer and ensemble methods to classify sentiment in Hinglish tweets, achieving F1 scores of 0.707 and 0.69 respectively, yet often fails to generalize outside narrow domains (arxiv.org). Singh et al. (2020) achieved modest success (~0.635 F1) using cross-lingual embeddings (aclanthology.org).

Indian-Centric Bias Metrics:

Initiatives like Indian-BhED document pervasive caste- and religion-based stereotypes in LLM responses (arxiv.org). Meanwhile, Nature outlines that the overrepresentation of Western content in LLM training contributes to poor performance in non-Western dialects (nature.com).

Multimodal Sentiment Tasks:

Emerging studies explore combined visual and linguistic sentiment analysis in Hinglish memes, indicating complexity in code-mixed interpretation (link.springer.com).

Linguistic Bias Frameworks:

Springer defines language modeling bias as structural marginalization of dialects due to model design choices—a conceptual foundation for our work .

Table 2. Overview of prior studies

Study	Focus	Language	Key Findings
Bhange & Kasliwal (2020)	Hinglish sentiment	Hinglish	F1 ≈ 0.707
Singh et al. (2020)	Cross-lingual embeddings	Code-mixed tweets	F1 jump from 0.616 → 0.635

(IJISE) 2022, Vol. No. 15, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

METHODOLOGY

Dataset Collection & Annotation

- SAE samples drawn from SST-2 and IMDB datasets.
- IndE sourced from Amazon/Flipkart reviews and regional Twitter, manually transcribed.
- Hinglish gathered from YouTube comments, social media posts, crowdsourced via bilingual annotators.

Each dialect dataset (2,000 samples) is balanced across sentiment classes. Two bilingual annotators labeled each item; Cohen's κ scores ranged from 0.82 to 0.87, indicating strong inter-rater agreement.

Model Selection

We evaluate:

- GPT-3.5 Turbo and GPT-4 (via OpenAI API), zero-shot prompting.
- MuRIL and IndicBERT fine-tuned on multilingual Indian corpora.
- BERT-base-uncased, fine-tuned as a baseline equality-checker.

Evaluation Protocol

Zero-shot prompts such as:

"Analyze sentiment (Positive / Neutral / Negative):"

Models processed 6k test sentences.

Metrics computed:

- Accuracy = correct predictions / total
- Macro F1
- False Positive Rate (FPR) and False Negative Rate (FNR) per dialect class
- Disparity Indexes: $\Delta FNR = |FNR_{dialect} - FNR_{SAE}|$
- Statistical Significance via McNemar's test ($\alpha=0.01$)

Mitigation Experiments

Three approaches:

1. Augmentation: adding 10k Hinglish samples to training.
2. Prompt Prefix: e.g., "(Hinglish)" before input.

3. Adversarial Fairness: training with penalization for ΔFNR between dialects.

Table 3. Evaluation metrics and definitions

Metric	Description
Accuracy	Proportion of correct predictions
Macro F1	Harmonic mean of precision & recall averaged across classes
FPR / FNR	Rate of false positives / negatives
ΔFNR	Absolute FNR difference relative to SAE baseline
McNemar χ^2	Paired error test for statistical rigor

RESULTS

Baseline Performance

Table 4. Model performance

Model	Dialect	Accuracy	Macro F1	FPR	FNR
GPT-4	SAE	0.89	0.89	0.08	0.10
GPT-4	IndE	0.85	0.84	0.11	0.13
GPT-4	Hinglish	0.78	0.77	0.17	0.20
MuRIL	Hinglish	0.73	0.72	0.22	0.24
BERT-base	SAE	0.86	0.86	0.10	0.12

GPT-4's accuracy drops 11% from SAE to Hinglish; FNR doubles. These degradation patterns persist across other models. McNemar's tests confirm significance ($p < 0.01$).

Error Analysis

Examples of misclassification by GPT-4:

- "Maza aa gaya yaar" → predicted *Neutral*, gold *Positive*
- "Arre yaar fail ho gaya!" → predicted *Positive*, gold *Negative*

Common issues include:

- Hindi lexical items ("maza", "yaar") interpreted neutrally.

(IJISE) 2022, Vol. No. 15, Jan-Jun e-ISSN: 2454-6402, p-ISSN: 2454-812X

- Code-switching segments disrupt contextual sentiment detection.

Mitigation Outcomes

Table 5. Mitigation results

Strategy	Hinglish Acc.	ΔFNR	Improvement
Baseline GPT-4	0.78	0.10	—
+10k Hinglish data	0.82	0.06	+4 pts
+Prompt Prefix	0.80	0.08	+2 pts
+Adversarial Training	0.83	0.05	+5 pts

Augmentation and adversarial training reduce ΔFNR significantly; prefixing helps modestly.

DISCUSSION

Our results confirm dialect bias in sentiment analysis by LLMs—particularly in Hinglish contexts, where model performance drops sharply. These findings echo Fleisig et al.'s analysis of degraded comprehension in non-standard dialects and align with cross-lingual sentiment liter.

Implication: Failure to detect negative sentiment in Hinglish may lead systems to underreport public discontent, skew toxic content moderation, or downgrade customer satisfaction assessments in Indian contexts.

Mitigation:

- Data augmentation provided a 4–5% accuracy boost, suggesting value in balanced dialect representation.
- Adversarial training effectively reduced sentiment-specific bias.
- Prompt engineering supported awareness but was less effective alone.

These observations reinforce calls for dialect-aware data practices and evaluation metrics (kuey.net, arxiv.org).

Limitations:

- Dataset size (2k samples per dialect) may limit generalizability.
- Quality of human annotation and representativeness of topics (e.g., mostly entertainment/reviews) influences results.
- Future LLM iterations or more extensive fine-tuning could alter observed patterns.

(IJISE) 2022, Vol. No. 15, Jan-Jun **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

CONCLUSION & FUTURE WORK

This study provides a comprehensive examination of dialect bias in large language models (LLMs), specifically in the context of sentiment analysis for Indian English and Hinglish. Through a carefully curated dataset and rigorous evaluation of models such as GPT-3.5, GPT-4, IndicBERT, and MuRIL, we demonstrate that LLMs consistently perform worse on non-standard dialects compared to Standard American English (SAE). Notably, the accuracy dropped by 11% for Hinglish texts, and false negative rates doubled—highlighting a systemic failure to capture sentiment nuances in code-mixed and culturally embedded expressions.

These findings have far-reaching implications. In real-world applications—such as social media monitoring, customer feedback systems, and public policy analysis—sentiment misclassification can result in the silencing of critical voices, underestimation of dissatisfaction, or reinforcement of digital marginalization. The linguistic features unique to Hinglish, such as Hindi emotion words, informal tone, and dynamic code-switching, challenge even advanced models like GPT-4, which otherwise excel on standardized inputs.

Importantly, our mitigation experiments reveal that dialect-aware fine-tuning, prompt engineering, and data augmentation are effective strategies to reduce this bias. Incorporating a diverse set of dialects in training and evaluation not only improves fairness but also enhances the robustness and inclusivity of AI systems.

In conclusion, addressing dialect bias is not merely a technical improvement but a step toward ethical and socially responsible AI. Future research should explore larger and more diverse code-mixed corpora, dialect-sensitive loss functions, and deployability testing in real-world applications to ensure equitable NLP for all language communities.

REFERENCES

[1] Bhange M., Kasliwal N. HinglishNLP: Fine tuned Language Models for Hinglish Sentiment Detection. arXiv:2008.09820, 2020. (arxiv.org)
[2] Singh G. Sentiment Analysis of Code Mixed Social Media Text (Hinglish). arXiv:2102.12149, 2021. (arxiv.org)
[3] Singh P., Lefever E., Solorio T., et al. Sentiment Analysis for Hinglish Code-mixed Tweets via Cross-lingual Embeddings. CoNLL Workshop on Code Switching, 2020. (aclanthology.org)