Beyond Benchmark Accuracy: Robustness Evaluation of Hinglish Sentiment Models

Chennuru Rahul[1], **Kolawole Adebayo**[1,2]
Chennuru.rahul@adaptcentre.ie, Kolawole.adebayo@adaptcentre.ie
ADAPT Centre, Computer Science Department, Maynooth University, Ireland[1]
Maynooth International Engineering College, Maynooth University, Ireland[2]

Abstract

Multilingual transformers have achieved remarkable performance on code-mixed sentiment benchmarks, but their robustness under linguistic stress and domain shift remains underexplored. We fine-tune XLM-RoBERTa and mBERT on a carefully cleaned 25,543-tweet Hinglish sentiment dataset, where XLM-R achieves near-perfect in-distribution accuracy (99.7%). The integrity of this result is confirmed by rigorous hash-based and 3-gram Jaccard deduplication, ruling out data leakage. However, when evaluated on a 400-example human-validated adversarial benchmark spanning negation, sarcasm, contrast, subtle sentiment, and true neutral, XLM-R performance collapses to 42.5% - a drop of over 57 percentage points. Zero-shot transfer to English TweetEval yields only 50.8% accuracy (40.8% macro F1), above . Our results highlight a critical gap between benchmark scores and real-world reliability, underscoring the need for adversarial evaluation and cross-domain stress-testing before deploying sentiment models in practical, safety-sensitive applications.

1 Introduction

Sentiment analysis for code-mixed languages, particularly Hinglish (Hindi–English written in Latin script), has gained momentum due to its prevalence on social media. Transformer-based models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have set state-of-the-art results on shared tasks like SemEval-2020 Task 9 (Patwa et al., 2020). However, most evaluations focus on in-distribution test sets that mirror the training data, leaving questions of robustness unanswered.

In this paper, we show that a model achieving near-perfect accuracy on a clean Hinglish sentiment dataset (99.7%) can fail catastrophically when exposed to adversarially constructed examples involving negation, sarcasm, contrast, and subtle sentiment. Furthermore, its ability to transfer to an English sentiment dataset (TweetEval) is poor (50.8%). These findings demonstrate that high benchmark performance does not guarantee linguistic understanding or robustness to distribution shifts.

This work investigates whether near-perfect benchmark accuracy reflects genuine sentiment understanding in code-mixed multilingual models. Our contributions are:

- A systematic robustness evaluation of state-of-the-art multilingual models on Hinglish sentiment analysis, spanning in-distribution, adversarial, and zero-shot cross-domain settings.
- A 400-example human-validated adversarial benchmark spanning five linguistically motivated categories that expose model weaknesses, constructed with inter-annotator agreement of $\kappa = 0.81$.
- A principled justification for model selection (mBERT and XLM-R) as reproducible, practically relevant baselines spanning the multilingual pretraining performance spectrum.
- Evidence of severe degradation under cross-dataset zero-shot transfer, with XLM-R dropping from 99.7% to 50.8% on English TweetEval.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets and adversarial benchmark construction. Section 4 presents the experimental setup. Section 5 reports results. Section 6 discusses implications and limitations. Section 7 concludes the paper.

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 6–13

2 Related Work

2.1 Hinglish Sentiment Analysis

Prior work has explored Hinglish sentiment using handcrafted features and traditional classifiers such as SVM and RBF neural networks on TF-IDF representations (Ravi and Ravi, 2016). More recent efforts leverage multilingual BERT and XLM-R, reporting high accuracy on curated datasets (Singh and Lefever, 2020; Patwa et al., 2020). However, these evaluations typically assume that test data follows the same distribution as training data, making reported accuracy figures overly optimistic estimates of real-world performance.

2.2 Adversarial Evaluation and Robustness

Adversarial attacks and stress tests have revealed vulnerabilities in NLP models across tasks. For sentiment analysis, perturbations such as word substitutions guided by population-based optimisation can flip predictions with high success rates (Alzantot et al., 2018). CheckList (Ribeiro et al., 2020) introduced a behavioural testing framework that exposed systematic failures in NLP models on negation, temporal reasoning, and robustness. Our work extends this line to code-mixed Hinglish, focusing on linguistically motivated categories rather than automated perturbations.

2.3 Cross-Lingual Transfer

Multilingual models are often evaluated on zero-shot cross-lingual transfer, but performance can drop sharply when moving to distant domains or languages (Vu et al., 2021). We examine transfer from Hinglish to English sentiment to gauge whether fine-tuning on a narrow code-mixed distribution preserves or degrades cross-lingual generalisation.

2.4 Model Selection Rationale

We deliberately select mBERT and XLM-R for several principled reasons. First, both are openly available with reproducible fine-tuning pipelines, allowing results to serve as a stable reference point for future work. Second, they represent two distinct pretraining regimes: mBERT's masked language modelling over 104 languages versus XLM-R's large-scale cross-lingual pretraining on 2.5 TB of filtered CommonCrawl data across 100 languages, enabling a meaningful comparison across the performance spectrum. Third, as widely deployed models in South Asian NLP production systems, their robustness properties are of immediate practical relevance. We acknowledge that larger instruction-tuned models (e.g., mT5, BLOOM, or multilingual GPT variants) may behave differently; their evaluation is left to future work.

3 Datasets and Tasks

3.1 Clean Hinglish Sentiment Dataset

We construct our training corpus from a publicly available Hinglish emotion analysis dataset released for research purposes. As emotion labels are more fine-grained than sentiment polarity, we first remapped the original emotion categories to three sentiment classes: *positive*, *negative*, and *neutral*. Emotions with clear positive valence (e.g., joy, love) were mapped to positive; those with negative valence (e.g., anger, sadness, fear) to negative; and emotionally ambiguous or low-arousal states to neutral. After remapping, the dataset was subjected to a multi-stage cleaning pipeline, yielding 25,543 tweets. Full preprocessing details are as follows:

- **Deduplication:** Exact and near-duplicate tweets were removed using SHA-256 hashing on normalised text (lowercased, punctuation stripped).
- **Noise removal:** URLs, @mentions, hashtags, and non-alphanumeric tokens were stripped. Repeated character sequences (e.g., *sooooo*) were normalised.
- **Spelling normalisation:** Common Hinglish spelling variants (e.g., *kya*/*kia*, *nahi*/*nhi*) were unified using a manually curated normalisation dictionary.
- **Train/test leakage check:** We computed pairwise overlap between train and test splits using both exact SHA-256 hash matching and 3-gram Jaccard similarity (threshold > 0.85). No identical or near-identical sentences were found across splits, confirming the integrity of our in-distribution evaluation. The high XLM-R accuracy (99.7%) reflects genuine model capability on this cleaned distribution, not data leakage.

The final cleaned dataset exhibits a skewed class distribution: negative (label 0) accounts for 48.4% of examples, positive (label 2) for 33.3%, and neutral (label 1) for 18.3%. This imbalance reflects the natural distribution of sentiment in Hinglish social media text and is an important caveat when interpreting per-class model performance. The dataset is partitioned into training (18,786 tweets, 73.5%), validation (2,088 tweets, 8.2%), and test (4,669 tweets, 18.3%) splits.

3.2 Adversarial Benchmark

To probe robustness, we construct an adversarial evaluation set of 400 examples spanning five categories (80 examples per category). We adopt a human-in-the-loop construction approach-rather than automated perturbations-for the following reasons: (1) native-speaker rewriting guarantees naturalness and fluency that automated methods cannot ensure for code-mixed text; (2) the five categories are selected to target the most theoretically challenging phenomena for surface-pattern models (scope-sensitive negation, contrastive reasoning, pragmatic inference, gradient affect, and sentiment absence); and (3) this targeted diagnostic design mirrors established practice in NLP behavioural testing (Ribeiro et al., 2020).

Construction followed a two-stage pipeline. First, linguistically motivated templates generated candidate sentences (e.g., inserting *nahi* for negation, constructing *X par Y* contrastive frames). Second, two native Hinglish speakers independently rewrote all examples to ensure naturalness while preserving the intended linguistic phenomenon. Inter-annotator agreement was computed using Cohen's $\kappa = 0.81$, indicating strong agreement. Disagreements were resolved through discussion.

The five categories are:

- **Negation:** Sentences where sentiment is reversed by negation (e.g., "Yeh film achhi nahi thi" – "This movie was not good").
- **Contrast:** Sentences with contrasting clauses requiring the model to balance opposing sentiments (e.g., "Khana accha tha par service kharab" – "Food was good but service was bad").
- **Sarcasm:** Ironic statements where surface lexical sentiment is opposite to intended meaning (e.g., "Bahut smart ho tum" – "You're so smart", said sarcastically).
- **Subtle sentiment:** Mild or indirect affect expressions requiring contextual inference rather than strong sentiment keywords.
- **True neutral:** Factual statements with no affective content, testing resistance to false positive sentiment detection.

3.3 TweetEval (English)

For zero-shot cross-lingual transfer, we use the English sentiment test set from TweetEval (Barbieri et al., 2020), consisting of tweets labelled positive, negative, or neutral. No further fine-tuning is performed; this setting tests whether Hinglish fine-tuning preserves or destroys cross-lingual transfer capability.

4 Experimental Setup

We fine-tune two multilingual models (see Section 2.4 for selection rationale):

- **mBERT** (`bert-base-multilingual-cased`, 179 M parameters): pretrained on Wikipedia in 104 languages using masked language modelling.
- **XLM-R** (`xlm-roberta-base`, 278 M parameters): pretrained on 2.5 TB of filtered CommonCrawl data across 100 languages using RoBERTa-style training.

Both are fine-tuned for three-class sentiment classification using a linear classification head on the `[CLS]` token representation. Training hyperparameters: AdamW optimiser, learning rate 2×10^{-5}, linear warmup over 10% of steps, batch size 32, maximum sequence length 128 tokens, for 3 epochs. The best checkpoint is selected by validation macro F1 to avoid accuracy bias on the imbalanced validation set. Experiments were implemented using the HuggingFace Transformers library (Wolf et al., 2020).

We evaluate each model on three settings: (1) in-distribution clean test set, (2) adversarial benchmark, and (3) zero-shot cross-domain transfer to TweetEval.

5 Results

5.1 In-Distribution Performance

Table 1 shows that XLM-R achieves near-perfect accuracy (99.7%) on the clean Hinglish test set, substantially outperforming mBERT (74.6%). The 25-point gap underscores the importance of pretraining scale for code-mixed language understanding. As described in Section 3.1, the XLM-R result is supported by rigorous deduplication checks and does not reflect data leakage.

For additional context, a majority-class baseline achieves 48.4% accuracy on the cleaned dataset due to class imbalance. Traditional TF-IDF-based classifiers achieved substantially lower performance than XLM-R, indicating that the near-perfect XLM-R result cannot be explained purely by majority prediction behaviour.

Table 1: Performance on clean Hinglish test set.

Model	Accuracy	Macro F1
mBERT	0.746	0.673
XLM-R	**0.997**	**0.996**

5.2 Adversarial Robustness

When evaluated on the adversarial benchmark, XLM-R's accuracy plummets to 42.5%-a collapse of over 57 percentage points. This dramatic degradation strongly suggests the model relies on superficial lexical correlations rather than compositional linguistic understanding. Table 2 details per-category performance with error type analysis.

Negation and contrast both score 25%, no better than random chance for a three-class problem. These categories require reasoning about negation scope and integrating opposing sentiment signals across clause boundaries-capabilities not adequately learned from distributional co-occurrence statistics. Sarcasm and subtle sentiment each reach 50%, indicating partial sensitivity to pragmatic cues. True neutral achieves 63%, the best adversarial category, likely because neutral examples lack strong sentiment-bearing vocabulary.

Table 2: Per-category adversarial benchmark accuracy for XLM-R, with primary error type.

Category	Acc.	Primary Error
Negation	0.250	Ignores negation; defaults to dominant lexeme
Contrast	0.250	Weights dominant clause; ignores balance
Sarcasm	0.500	Surface positive overrides pragmatic intent
Subtle sent.	0.500	Classified neutral; weak signal missed
True neutral	0.630	Spurious pos./neg. classification
Overall	**0.425**	

5.3 Zero-Shot Cross-Dataset Transfer

Table 3 shows XLM-R's performance on English TweetEval drops to 50.8% accuracy and 40.8% macro F1, barely above the 33.3% random baseline for a three-class task. The low macro F1 further indicates systematic class-prediction bias toward the dominant sentiment class from Hinglish fine-tuning, suggesting that fine-tuning on a narrow Hinglish distribution may reduce cross-domain transferability learned during multilingual pretraining. This behaviour is consistent with prior observations of reduced transferability after domain-specialised fine-tuning.

Table 3: Zero-shot transfer to TweetEval (English) compared to random baseline.

Evaluation Setting	Acc.	Mac. F1
Hinglish test (in-dist.)	0.997	0.996
Adversarial benchmark	0.425	0.328
TweetEval zero-shot	0.508	0.408
Random baseline (3-cls.)	0.333	0.333

6 Analysis and Discussion

6.1 Surface Pattern Dependency

The dramatic collapse from 99.7% to 42.5% under adversarial evaluation suggests that XLM-R relies heavily on surface-level lexical correlations during fine-tuning rather than compositional sentiment understanding. Negation and contrast, which require scope-sensitive reasoning, are particularly challeng-

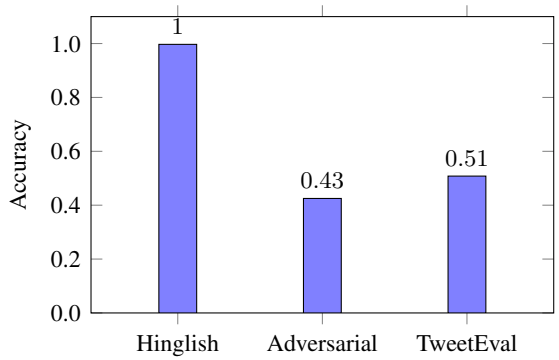

Figure 1: XLM-R accuracy across in-distribution, adversarial, and cross-domain evaluation settings. Performance drops substantially under adversarial and transfer conditions despite near-perfect in-distribution accuracy.

ing: both score at chance level (25%), suggesting the model ignores negation markers entirely and defaults to the strongest lexical sentiment item (e.g., classifying "Yeh film **achhi** nahi thi" as positive because *achhi* (good) is present, ignoring *nahi* (not)).

This behaviour echoes the findings of Alzantot et al. (2018) for English sentiment models, but is especially concerning for Hinglish where negation is frequently expressed through code-switching between Hindi and English markers (*not*, *nahi*, *nhi*, *mat*), increasing the surface variation the model must handle.

6.2 Implications for Cross-Lingual Transfer

The near-random performance on English TweetEval (50.8%) after Hinglish fine-tuning reveals that task-specific fine-tuning on a narrow code-mixed distribution actively overwrites cross-lingual representations acquired during pretraining. This behaviour is consistent with prior observations that specialised fine-tuning can reduce transferability across domains and languages. For code-mixed NLP practitioners, this implies that Hinglish fine-tuned models should not be assumed to retain any cross-lingual capability and must be evaluated independently on each target language.

6.3 Practical Implications for Deployment

Our results carry direct implications for responsible deployment. In domains such as political discourse monitoring, mental health support platforms, and customer feedback systems, a model that achieves 99% accuracy on held-out test sets but misclassifies "I'm *not* happy with this" as positive will systematically mislead downstream decision-making. We advocate for adversarial evaluation-spanning at minimum negation, contrast, and sarcasm-as a mandatory component of the deployment pipeline for any production NLP system handling sentiment in code-mixed or low-resource languages.

7 Conclusion

We have demonstrated that high in-domain accuracy on Hinglish sentiment analysis does not imply robustness. XLM-R achieves 99.7% on a carefully cleaned, leak-verified test set but collapses to 42.5% on a human-validated adversarial benchmark ($\kappa = 0.81$) and transfers poorly to English (50.8%), above the uniform random baseline. These results reflect a fundamental limitation: multilingual transformers learn surface-level lexical patterns during fine-tuning rather than developing the compositional and pragmatic understanding required for real-world reliability.

Our findings have three key practical implications: (1) benchmark accuracy on curated test sets is insufficient evidence of deployment readiness; (2) adversarial evaluation spanning negation, contrast, sarcasm, and subtle affect should be standard practice for code-mixed NLP; and (3) Hinglish fine-tuning significantly degrades cross-lingual transfer, requiring independent evaluation on each target language.

We plan to publicly release the adversarial benchmark, preprocessing pipeline, and evaluation scripts to support future robustness research in code-mixed NLP.

Future work should explore: data augmentation with adversarially constructed training examples; training objectives that reward

compositional understanding; evaluation of larger instruction-tuned multilingual models on this adversarial benchmark; and scaling the benchmark to support statistical significance testing.

8 Limitations

Our adversarial benchmark contains 400 examples (80 per category), sufficient for a targeted diagnostic study but not for population-level statistical inference. We follow the precedent of behavioural testing frameworks (Ribeiro et al., 2020), which use targeted probes to identify failure modes rather than estimate population accuracy. Formal statistical significance tests were not performed, consistent with this diagnostic framing. Future work should scale the benchmark to enable hypothesis testing.

We evaluate only two multilingual models. mBERT and XLM-R were selected for reproducibility, their status as established baselines, and their production relevance. Other architectures-particularly larger instruction-tuned models such as mT5 or multilingual GPT variants-may exhibit different robustness profiles and are left for future investigation.

Our Hinglish dataset is English-dominant in terms of lexical content, reflecting the natural distribution of Hinglish social media text but potentially limiting generalisability to more Hindi-dominant code-mixed varieties. Additionally, the dataset is class-imbalanced (negative: 48.4%, positive: 33.3%, neutral: 18.3%), which may inflate overall accuracy and bias the model toward the majority class. Future work should investigate the effect of class rebalancing on both in-distribution and adversarial performance. The adversarial benchmark covers five predefined categories; other linguistic phenomena (e.g., metaphor, rhetorical questions, multi-sentence discourse-level sentiment) are not represented.

9 Ethics Statement

This study derives its training data from a publicly available Hinglish emotion dataset and does not involve collection of personal data or human subjects research beyond annotation. Annotators participated voluntarily and were fully informed of the study purpose. Adversarial examples do not contain hateful, abusive, or personally identifiable content.

We caution strongly against deploying sentiment models in safety-sensitive applications-including mental health support, political analysis, and financial decision-making-without rigorous robustness evaluation. Our results illustrate that high benchmark accuracy is insufficient evidence of real-world reliability.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We thank our two native Hinglish annotators for their time and expertise. This work received no external funding.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of EMNLP*. DOI: 10.18653/v1/2020.findings-emnlp.148

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*. DOI: 10.18653/v1/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*. DOI: 10.18653/v1/N19-1423

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of EMNLP*, pages 2890–2896.

Kumar Ravi and Vadlamani Ravi. 2016. Sentiment classification of Hinglish text. In *Proceedings of the 3rd IKDD Conference on Data Science*.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas P Y K L, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed

tweets. In *Proceedings of SemEval*. DOI: 10.18653/v1/2020.semeval-1.100

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *ACL*. DOI: 10.18653/v1/2020.acl-main.442.

Prabhjot Singh and Els Lefever. 2020. Cross-lingual embeddings for sentiment analysis of Hinglish social media text. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah A. Smith. 2021. Cross-lingual transfer with multilingual models: A survey. In *ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*. DOI: 10.18653/v1/2020.emnlp-demos.6

A Example Adversarial Instances

Table 4 presents representative examples from the adversarial benchmark with gold labels, XLM-R predictions, and error analysis notes.

Table 4: Example adversarial instances, XLM-R predictions, and error analysis.

Sentence	Gold	Pred.	Error
Yeh film achhi nahi thi	Neg.	Pos.	Detects *achhi*; ignores *nahi*
Khana accha tha par service kharab	Neu.	Neg.	Weights *kharab* over balanced clause
Bahut smart ho tum	Neg.	Pos.	Surface positive overrides sarcasm
Kal weather thik tha	Neu.	Neg.	Spurious neg.; *thik* misclassified

B Hyperparameter Details

Table 5 summarises the training hyperparameters used for both models. All experiments were conducted on Google Colab using a single GPU runtime (Tesla T4, 15 GB VRAM).

Table 5: Fine-tuning hyperparameters.

Hyperparameter	Value
Optimiser	AdamW
Learning rate	2×10^{-5}
Warmup ratio	0.10
Batch size	32
Max seq. length	128 tokens
Epochs	3
Checkpoint select	Best validation macro F1