SarcasmLens: Detecting Sarcasm in Hindi–English Code-Mixed Text

Akshat Kadia
202203029
GitHub Repository

Abstract

Detecting sarcasm in social media text is a difficult task because the intended meaning is often the opposite of what is written. This challenge becomes even greater in Hindi–English code-mixed content, where users freely switch languages and rely on cultural context. In this work, I present *SarcasmLens*, a two-stage sarcasm detection system. First, I develop a strong baseline using TF–IDF features and simple linguistic cues. While this baseline achieves high accuracy, data analysis reveals that parts of the dataset contain repeated and template-based patterns that inflate performance. To address this, I extend the system using a hybrid ensemble approach that combines a multilingual transformer (XLM-RoBERTa) with traditional machine learning models. The final system captures both surface-level sarcasm cues and deeper contextual meaning, resulting in stronger and more reliable predictions.

1 Introduction

Sarcasm is a form of expression where the emotional intent does not match the literal meaning of the words [6]. For example, phrases like *"Great, another meeting!"* use positive language to convey frustration. Detecting sarcasm automatically is difficult, especially on social media where informal writing, emojis, and cultural references are common.

The problem is even harder for Hindi–English code-mixed text. Users often switch languages within a sentence, use Romanized Hindi, and rely on shared social and political context. Code-mixed language processing for Indian languages has been studied in several shared tasks and datasets [2, 8], but sarcasm detection in this setting is still challenging. This work aims to build a sarcasm detection system that works well for such code-mixed content while remaining transparent and easy to interpret.

2 Related Work

Early sarcasm detection approaches focused on surface features such as punctuation, capitalization, and sentiment words [6]. With the introduction of transformer-based models such as BERT [4], more recent work has shown that contextual embeddings can capture more subtle patterns in text.

However, most existing systems are designed for monolingual English data. For code-mixed Indian languages, several datasets and shared tasks have been proposed, particularly for sentiment and sarcasm detection [10, 8, 2]. At the same time, prior studies have highlighted issues of data leakage and annotation quality in NLP datasets [7, 5], which motivates careful analysis of the dataset used in this work.

3 Dataset and Data Quality

The dataset used in this study contains 11,919 Hindi–English social media comments labeled as sarcastic or non-sarcastic. After removing exact duplicates, 11,854 samples remained.

Data Quality Observations

During analysis, several issues were identified:

- Some comments are repeated or nearly identical
- A few texts appear in both training and test sets
- Many sarcastic examples follow fixed templates

Such issues are similar to data leakage and design problems reported in other NLP datasets [7]. These problems can make models appear more accurate than they really are, so care was taken when interpreting results. Following the idea of "datasheets for datasets" [5], this work explicitly documents these limitations.

4 Baseline System

As a starting point, I built a baseline model using TF–IDF features and simple linguistic indicators.

Preprocessing

The text was cleaned using the following steps:

- Removal of URLs and user mentions
- Conversion of emojis into text
- Splitting of hashtags into words
- Normalization of elongated words
- Preservation of punctuation such as ! and ?

Features and Models

The baseline features include TF–IDF word and phrase statistics along with punctuation counts, text length, capitalization patterns, and sentiment scores. Several standard classifiers were tested, including Logistic Regression, Linear SVM, and Random Forest.

5 Hybrid Ensemble Approach

5.1 Why Use a Transformer?

Although the baseline models performed well, they rely heavily on repeated patterns in the dataset and struggle to understand meaning beyond surface-level cues. To overcome this, I incorporated XLM-RoBERTa, a multilingual transformer model trained on large-scale cross-lingual data [3]. It can naturally process code-mixed text without language detection or translation and follows the broader trend of using multilingual transformers for cross-lingual transfer [9]. For a more intuitive explanation of XLM-RoBERTa and its usage, I also refer to the practical overview in [1].

5.2 Ensemble Design

The final prediction is produced by combining the probabilities from traditional models and XLM-RoBERTa [3, 1]:

$$P_{\text{final}} = \alpha P_{\text{traditional}} + (1 - \alpha) P_{\text{XLM-R}}.$$

The best value of α was found to be 0.4, giving slightly more weight to the transformer predictions.

6 Experimental Setup

- 80–20 train-test split
- 5-fold stratified cross-validation
- Evaluation metrics: F1-score, Accuracy, and AUC-ROC

7 Results

Model	Accuracy	F1-score	AUC
Linear SVM	0.9746	0.9746	0.9951
XLM-RoBERTa	0.9812	0.9811	0.9978
Ensemble	**0.9860**	**0.9859**	**0.9989**

Table 1: Performance comparison of different models

The ensemble model performs best across all metrics, especially on texts that are ambiguous or contain mixed-language sarcasm.

8 Error Analysis

Manual inspection showed that baseline models often rely on known sarcastic phrases and punctuation. The transformer model is better at understanding context, especially in political and cultural examples, which aligns with the strengths of contextual and multilingual models reported in earlier work [6, 3]. Errors usually occur when sarcasm depends on external knowledge or conversation history that is not present in the text.

9 Limitations

- Dataset contains repeated and synthetic patterns
- No conversational context is available
- Results may not generalize to other datasets

These limitations are consistent with broader concerns about data quality and evaluation in NLP datasets [7, 5].

10 Conclusion and Future Work

This work shows that combining traditional linguistic features with multilingual transformer models leads to better sarcasm detection for code-mixed text. While high performance is achieved, the results should be viewed carefully due to dataset limitations. In the future, I plan to collect more realistic data, include conversation-level context, and extend this approach to other multilingual sarcasm detection tasks [10, 8].

References

[1] Aman Anand. *Cross Lingual Models (XLM-R): A deep dive into XLM-R*. 2020. URL: `https://medium.com/@aman.anand54321/cross-lingual-models-xlm-r-7d557302698b` (visited on 12/03/2025).

[2] Bharathi Raja Chakravarthi et al. “Overview of the Track on Sentiment Analysis for Code-Mixed Languages”. In: *FIRE Working Notes* (2020).

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of ACL*. 2020.

[4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL*. 2019.

[5] T. Gebru et al. “Datasheets for datasets”. In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[6] A. Joshi et al. “Sarcasm detection: An interdisciplinary approach”. In: *Journal of Artificial Intelligence Research* 58 (2017), pp. 165–207.

[7] A. Kapoor, S. Narayan, and R. R. Shah. “Data leakage in natural language processing: A survey and reproducibility assessment of the CodeMixed-IndianLangTweets sentiment corpus”. In: *Journal of Artificial Intelligence Research* 71 (2021), pp. 1275–1303.

[8] B. G. Patra et al. “Shared task on sentiment and sarcasm detection in code-mixed dialogue”. In: *Proceedings of the Forum for Information Retrieval Evaluation*. 2019, pp. 1–10.

[9] Jonas Pfeiffer et al. “MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer”. In: *arXiv preprint arXiv:2005.00052* (2020).

[10] S. Swami et al. “A corpus of English-Hindi code-mixed tweets for sarcasm detection”. In: *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. 2018, pp. 1–10.