Sarcasm Detection in Low Resource Indic Language Using Deep Learning Techniques

Akshat Jha akshatjh@usc.edu
Atharv Arya atharvar@usc.edu
Dheeraj Reddy dsreddy@usc.edu
Mudra Limbasia limbasia@usc.edu
Yamini Joshi yjoshi@usc.edu

Abstract

Sarcasm detection in low-resource Indic languages like Hindi and Bengali becomes a tedious task because of the lack of a large corpus of annotated data, unique linguistic structures, and rich cultural nuances of these languages that affect the expression of sarcasm. The data available in these languages include a lot of dialect variation, morphological complexity, and language/code-mixing making it difficult for the traditional methods to detect the presence of sarcasm in them. We aim to approach this problem by training and fine tuning context aware deep-learning techniques. These models capture the nuances of Indic Languages and make them better fit for this particular task. We experimented with mBERT, IndicBERT, and XLM-RoBERTa. XLM-RoBERTa out-performed the other two models by achieving an accuracy of 93.24%.

1 Introduction

Detecting sarcasm is one of the toughest challenges in natural language processing (NLP). It's essential for machines to understand sarcasm better if we want to improve human-machine interactions. The challenge lies in sarcasm's reliance on subtle cues like tone, context, and cultural knowledge—things that are often missing or ambiguous in written text. This problem becomes even more complex when dealing with low-resource Indic languages, where the lack of labeled data, mixed languages, regional dialects, and cultural diversity further complicate the task.

Our research tackles this challenge by leveraging advancements in deep learning and multilingual NLP. We began by collecting a robust sarcasm dataset from public platforms like Reddit and Twitter, where sarcasm thrives. The data underwent meticulous preprocessing to ensure it retained important nuances, such as dialect-specific phrases and mixed-language expressions, reflecting the rich variety of sarcasm in Indic languages.

At the core of our approach are powerful deep learning models—particularly transformer-based architectures and Large Language Models(LLMs). We focused on Multilingual BERT (mBERT)(Devlin, 2018), a model pre-trained on 104 languages, making it ideal for capturing both context and meaning across multiple languages. To further enhance performance, we fine-tuned IndicBERT (Dabre et al., 2021), a model specifically designed for Indic languages, to address challenges like data scarcity, morphological complexity, and cultural subtleties. But, we noticed that XLM-RoBERTa (Conneau, 2019), a Large Language Model(LLM) performed the best, and was able to detect the presence of sarcasm more effectively than the other two transformer-based models.

We evaluated our models using standard metrics such as ccuracy, Macro F1 Score, and recall, and analyzed their effectiveness using confusion matrices. Through this rigorous testing, our models demonstrated strong performance, with mBERT achieving an Accuracy of 81.01% and IndicBERT pushing this further to 83.81%. However, XLM-RoBERTa performed the best achieving an accuracy of 93.24%. These results highlight the models' ability to effectively generalize across the diverse linguistic and cultural contexts of Indic languages.

2 Related Works

Most of the recent progress in sarcasm detection has focused on high-resource languages like English. This is largely thanks to the availability of massive datasets and pre-trained models, which make it easier to adopt advanced deep learning techniques. For instance, Tan et al. (2023) showed that combining sentiment analysis with sarcasm detection using deep multi-task learning can significantly boost performance. Their findings highlight that sharing representations across related tasks helps models grasp the subtle nuances of sarcasm

better.

Another innovative approach comes from the work of Liu et al. (2024), who introduced the Image-Text Fusion Transformer Network. This model blends textual and visual information, effectively bridging the gap between literal meanings and sarcastic intent. By analyzing multimodal data—such as text paired with images—this method proves especially valuable for sarcasm expressed through a mix of visual and textual cues.

Multimodal techniques hold a lot of promise because sarcasm often relies on tone, facial expressions, or visual contrasts—things that simple text-based models struggle to capture. However, these approaches require large annotated datasets and significant computing power, both of which remain scarce for low-resource languages like Hindi, Bengali, and other Indic languages. This shortage limits the direct use of advanced techniques like CNNs, RNNs, and transformers in these environments.

Despite these challenges, some notable progress has been made for Indic languages. For example, Rathore et al. (2024) tackled the morphological richness of Hindi by focusing on sarcasm detection in tweets. They showed that task-specific features—like hashtags and slang—play a critical role in improving accuracy, particularly in low-resource settings where data is sparse.

Similarly, Pandey et al. (2024) proposed a hybrid Convolutional Neural Network (CNN-H) model to handle the complexities of multilingual social media content. By combining character-level and word-level embeddings, the CNN-H approach effectively addresses the challenges of code-mixing, where multiple languages like Hindi-English or Bengali-English are blended in a single sentence. This hybrid design ensures that both language and cultural nuances of sarcasm are captured more accurately, even with smaller datasets.

One of the most promising advancements in this space is Multilingual BERT (mBERT), introduced by Devlin (2018). Pre-trained on 104 languages, including several Indic ones, mBERT has demonstrated a strong ability to generalize across languages, even without explicit cross-linguistic alignment. By fine-tuning mBERT on domain-specific datasets, researchers can leverage its pre-trained knowledge to improve sarcasm detection, even when data is limited. Unlike traditional models, which often fail in multilingual settings, mBERT's flexibility offers a robust alternative.

To summarize, while sarcasm detection has seen significant breakthroughs in English, low-resource Indic languages continue to face hurdles like a lack of annotated data, code-mixed content, and vast linguistic diversity. Recent work points to promising directions, such as transformer-based models like mBERT, hybrid neural networks, and multimodal approaches. Expanding research in this area is critical to advancing NLP applications such as sentiment analysis, content moderation, and social media monitoring for Indic languages.

Our project aims to contribute to this effort by compiling diverse datasets from multiple sources and training models like mBERT, IndicBERT, and XLM-RoBERTa to accurately detect sarcasm in text.

3 Methodology

We gathered data from several sources to create a representative and varied dataset in order to create an efficient sarcasm detection framework. A number of pre-processing techniques were applied to the raw data in order to normalize the data. Our main classifiers were state-of-the-art deep learning models, such as mBERT, IndicBERT, and XLM-RoBERTa, which were trained using this pre-processed data. 1 depicts the general approach used in this investigation.

3.1 Datasets

We combined three datasets: the Hindi Tweets Dataset for Sarcasm Detection from Kaggle, Hindi tweets we scraped from Reddit, and the Bangla SARC(Bengali) dataset(Apon et al., 2022) from Kaggle, which includes Bengali tweets. The Hindi Kaggle dataset has nearly 16,000 tweets, while the Reddit dataset has 1,862 samples, and the Bengali dataset has 5,112. This adds up to 22,974 samples, but only 12,735 of these were unique.

Since we noticed an imbalance between the classes and we used the RandomOverSampler (Lauron and Pabico, 2016) to deal with it. It is a class provided by the imbalanced-learn library, it oversamples the minority class to upsample it and balance the dataset. The imbalance in dataset causes the model to favor the majority class, as it dominates the training data, leading to poor performance on the minority class. RandomOverSampler duplicates samples from the minority class till a balanced distribution is achieved. This is done to ensure that both the classes get equal representation

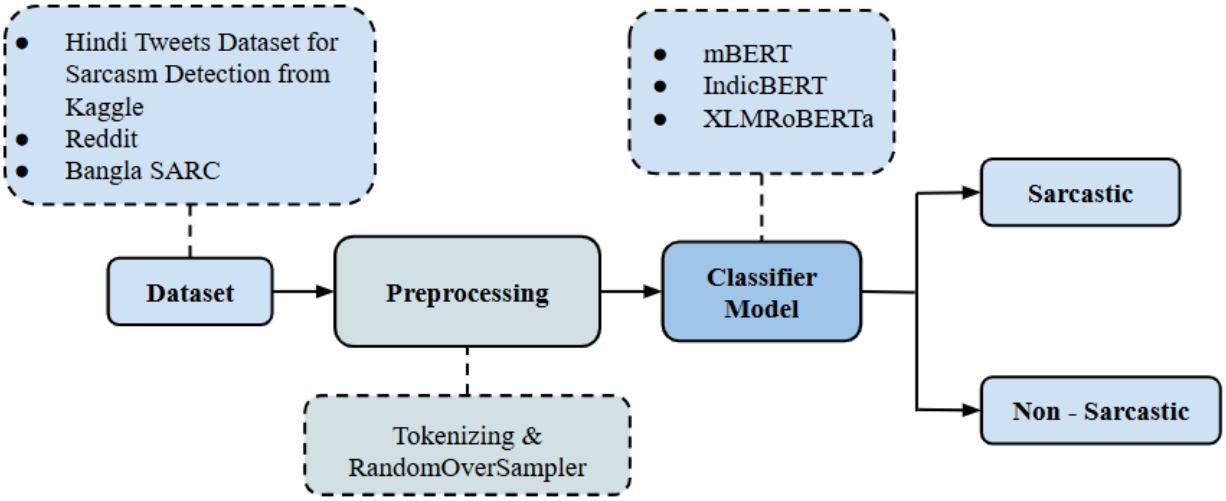

Figure 1: Proposed methodology

while training the model, improving its ability to make fair and accurate predictions.

3.2 Pre-processing

The compiled dataset underwent comprehensive pre-processing to prepare before being used to train the classifier models. The duplicate entries were eliminated from the dataset ensuring that we do not have any redundant data that could potentially skew the results.

After removing the duplicates the pre-processing pipeline that followed included cleaning and standardizing the text. Regular expressions were used to remove URLs, mentions (e.g., '@user'), hashtags, and extra spaces. To preserve the semantic meaning of emojis they were converted into descriptive texts. Removal of punctuations and stop words were also performed. The stop words that were removed were mainly from English, Hindi, and Bengali languages, this multi-lingual stopword set ensured the exclusion of unnecessary tokens.

The pre-processed and cleaned datasets were then merged into a single unified dataset with consistent labeling. This rigorous pre-processing ensured data uniformity while retaining key contextual and emotional queues for the task of detecting sarcasm.

3.3 Classifier Model

We fine tuned three pre-trained transformer models - **mBERT, IndicBERT, and XLM-Roberta** on the pre-processed and augmented dataset. The dataset was split into 72% training, 8% testing, and 20% validation. We used accuracy as our evaluation metric, which is calculated by comparing the predictions made by the model and the ground-truth labels.

- **Multilingual BERT** : mBERT (Multilingual BERT) is a multilingual version of BERT, pretrained on the Wikipedia corpora of 104 languages using a shared vocabulary, it facilitates cross-lingual comprehension and transfer learning for a variety of natural language processing (NLP) tasks.

  We tokenized the pre-processed data using mBERT tokenizer and fed into the mBERT classifier model. A learning rate of 1e-5, weight decay of 0.1, a batch size of 8, and 20 epochs were used. An EarlyStoppingCallback (patience = 3) was applied to halt training if validation loss did not improve. Logging occurred every 10 steps, and only the 2 best checkpoints were saved.

- **IndicBERT** : IndicBERT is a multilingual language model designed specifically for Indian languages. IndicBERT provides reliable and effective performance on a range of natural language processing applications in low-resource Indic language environments.

  We tokenized the model and trained it using ai4bharat/indic-bert tokenizer and model. Training parameters included a learning rate of 1e-6, weight decay of 0.1, a batch size of 8, and 20 epochs. Early stopping (patience = 3), logging every 10 steps, and limiting saved checkpoints to 2 were applied, similar to mBERT.

Model	Accuracy	Macro-F1 Score
mBERT	81.01%	0.77
IndicBERT	83.81%	0.77
XLM-RoBERTA	**93.24%**	**0.91**

Table 1: Results Obtained

- **XLM-RoBERTa** :XLM-RoBERTa is a multilingual variant of RoBERTa. Strong performance across a variety of multilingual NLP tasks is made possible by this Large Language Model, as it has been pre-trained on 100 languages utilizing self-supervised learning on vast volumes of text.

  The xlm-roberta-large tokenizer was used, with the XLM-Roberta model fine-tuned under a learning rate of 2e-6, weight decay of 0.01, a batch size of 8, and 20 epochs. An EarlyStoppingCallback with a higher patience value (8 evaluations) was applied. Logging and checkpoint-saving configurations remained consistent.

4 Results

We used accuracy as our evaluation metric, calculated by comparing the model's predictions with the ground-truth labels. Table 1 presents a summary of the results achieved by the three models, providing a clear comparison of their performance. Among the models evaluated, XLM-RoBERTA demonstrated the best performance, achieving an impressive accuracy of 93.24%. A detailed discussion of these results is provided below.

4.1 mBERT

The mBERT model, trained on the aforementioned dataset and parameters, achieved an accuracy of 81.01% with a Macro F1 score of 0.77. Figure 2 depicting the confusion matrix obtained, highlights that the model performs well on classifying the majority class (60.68%) but struggles slightly with the minority class, showing a misclassification rate of 16.07%. This indicates room for improvement in handling class imbalance.

The learning curve shown in Figure 3 reveals a consistent decrease in both training and validation loss over 20 epochs. However, a stability in the gap between training and validation loss is obseved in the later epochs, indicating that the model is robust and is not overfitting.

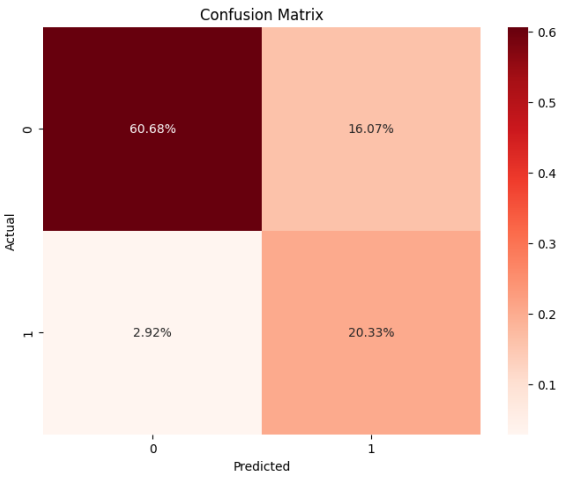

Figure 2: Confusion Matrix for the mBERT model

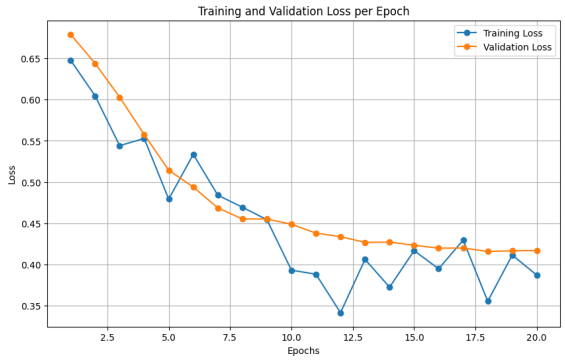

Figure 3: Learning Curve of the mBERT Model

4.2 IndicBERT

The IndicBERT model, trained using the specified dataset and parameters, performed well with an accuracy of 83.81% and a Macro F1 score of 0.77. As seen in the confusion matrix in Figure 4, the model successfully classifies the majority class, achieving a correct prediction rate of 69.63%. However, the minority class remains challenging, with a misclassification rate of 16.19%, highlighting the need for targeted strategies to address class imbalance and improve its predictive performance. The learning curve depicted in Figure 5 shows a steady decline in both training and validation losses over 12 epochs. The consistent downward trend suggests that the model is learning effectively, and the minimal gap between training and validation losses in the later stages indicates strong generalization without overfitting.Demonstrating that the training process was well-executed and optimized for the data.

In summary, the IndicBERT model performs admirably overall but has room for improvement when it comes to handling class imbalance. Incorporating techniques such as oversampling the

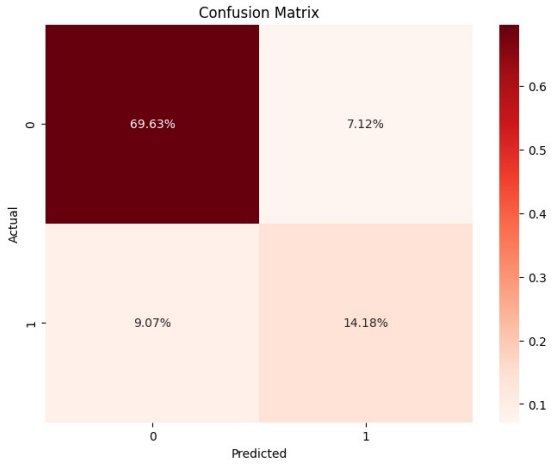

Figure 4: Confusion Matrix for the IndicBERT model

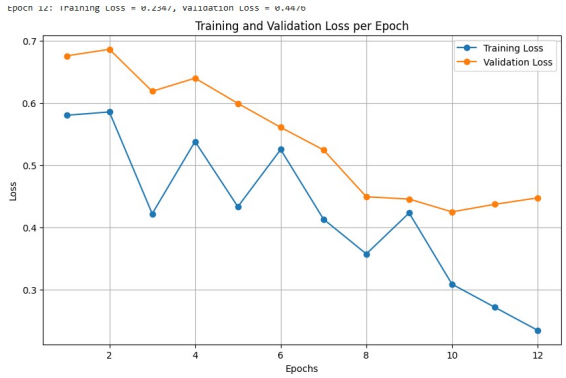

Figure 5: Learning Curve of the IndicBERT Model

minority class, class weighting, or introducing data augmentation could further refine the model's performance, particularly by enhancing recall and precision for the underrepresented class.

4.3 XLM-RoBERTa

The XLM-RoBERTa model achieved an accuracy of 93.24% with a Macro F1 score of 0.91. As shown in the confusion matrix in Figure 6, the model performs exceptionally well in classifying the majority class with a prediction accuracy of 71.64%. Similarly, the minority class also demonstrates strong performance, with only 1.64% misclassifications. This indicates the model's ability to handle class imbalance effectively, with minimal errors across both classes.

The learning curve in Figure 7 shows that the model converges very quickly during training. Initially, in the first few epochs the training loss drops sharply, suggesting that the data patterns are learnt by the model in an efficient manner. However, in the later epochs, the training loss continues to decrease, while the validation loss begins to increase. This trend highlights a point of overfitting, where the model trains intensively on the training data, losing its ability to generalize well on unseen data. In conclusion, the LLM demonstrates exceptional overall performance, achieving high accuracy and balanced results across classes.

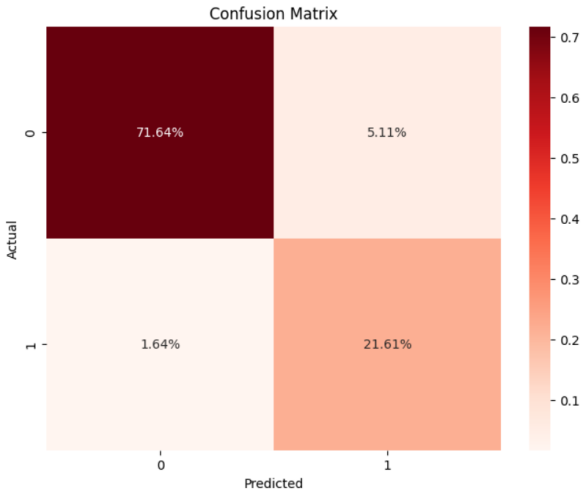

Figure 6: Confusion Matrix for the XLM-RoBERTa model

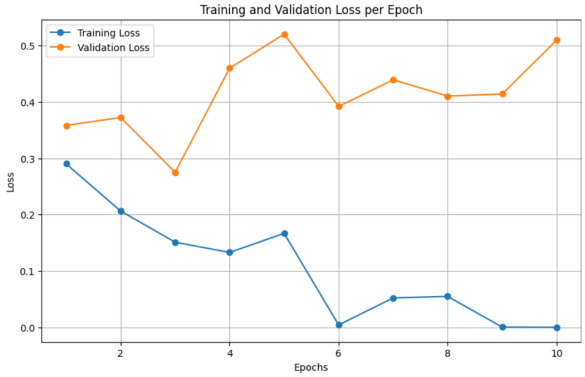

Figure 7: Learning Curve of the XLM-RoBERTa Model

5 Discussion

In this project, we tried to tackle the problem of sarcasm detection in low-resource Indic languages. We experimented with three deep learning models namely,- IndicBERT, mBERT, and XLM-RoBERTa, on a on a custom-labeled dataset containing texts in Hindi and Bengali languages.

We noticed that IndicBERT outperformed mBERT with an accuracy of 83.81%, and was able to efficiently capture the syntactic and semantic nuances of Indic languages due to its specialized pretraining on Indic scripts. mBERT, which was originally designed as a multilingual gener-

alis, struggled to capture the language specific nuances. However a Large Language Model(LLM), XLM-RoBERTa, making use of its robust multilingual representation and contextual embeddings, performed the best with an accuracy of 93.24%. Indicating the model robustness and capacity to generalise across languages and capture deeper linguistic patterns essential for sarcasm detection.

The Large language models (LLMs) demonstrated great promise, but we found that our model overfit during fine-tuning because of its quicker convergence. This was probably brought about by the dataset's small size and the model's high memorization ability. This can be addressed by reducing overfitting through the use of strategies like regularization, dropout, and early stopping. Furthermore, the model's capacity to generalize more effectively across unseen cases may be improved by data augmentation techniques or the acquisition of a larger dataset. Our study demonstrates the larger potential of large language models (LLMs) for sarcasm detection tasks, especially in languages with limited resources. When fine-tuned with pertinent datasets, LLMs exhibit significant adaptability to complex and culturally specialized tasks. They are a potential direction for future research because of their scalability and versatility, particularly when it comes to addressing NLP issues in underrepresented languages.

6 Future Work

Future research in sarcasm detection for low-resource Indic languages can concentrate on resolving existing issues and investigating more sophisticated techniques. Since larger and more varied datasets can greatly improve the generalizability of models, expanding datasets continues to be a top objective. In situations where labeled data is limited, methods like zero-shot and few-shot learning can be used to enhance performance and allow models to adjust to new languages or circumstances with little oversight. Understanding the subtleties of sarcasm in Indic languages, which are frequently linked to regional and social issues, would also require incorporating extra cultural and linguistic background.

Using multimodal techniques to supplement textual data is another exciting field. By capturing non-verbal clues like tone, facial expressions, and gestures—all of which are essential to sarcasm—combining text with visual, audio, and video data can increase detection accuracy. With its multilingual capabilities, sophisticated big language models like XLM-RoBERTa are ideal for managing intricate linguistic phenomena like code-mixing and the semantic nuances of Indic languages. Using websites like Facebook and YouTube, a pipeline for crowdsourced data collecting may be created to help with dataset development even more. High-quality annotations can be guaranteed by combining web scraping with heuristic-based pre-labeling and crowdsourcing evaluation. Additionally, exploring social graph-based sarcasm detection using Graph Neural Networks (GNNs) could capture user interactions, relationships, and conversational dynamics, offering a richer context for sarcasm detection in social media environments.

References

Tasnim Sakib Apon, Ramisa Anan, Elizabeth Antora Modhu, Arjun Suter, Ifrit Jamal Sneha, and MD Golam Rabiul Alam. 2022. Banglasarc: A dataset for sarcasm detection. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–5. IEEE.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maureen Lyndel C Lauron and Jaderick P Pabico. 2016. Improved sampling techniques for learning an imbalanced data set. *arXiv preprint arXiv:1601.04756*.

Jing Liu, Shengwei Tian, Long Yu, Xianwei Shi, and Fan Wang. 2024. Image-text fusion transformer network for sarcasm detection. *Multimedia Tools and Applications*, 83(14):41895–41909.

Rajnish Pandey, Abhinav Kumar, Jyoti Prakash Singh, and Sudhakar Tripathi. 2024. A hybrid convolutional neural network for sarcasm detection from multilingual social media posts. *Multimedia Tools and Applications*, pages 1–29.

Pramod Singh Rathore, Sachin Ahuja, Srinivasa Rao Burri, Ajay Khunteta, Anupam Baliyan, and Abhishek Kumar. 2024. *Deep Learning Techniques for Automation and Industrial Applications*. John Wiley & Sons.

Yik Yang Tan, Chee-Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. 2023. Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless personal communications*, 129(3):2213–2237.