

M – Sarcasm: A Robust Deep Learning Framework For Enhanced Sarcasm Detection Of Marathi Tweets

Pravin Patil^{1*} & Satish Kolhe²

^{1*}School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon.

² School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon

Abstract

Detecting sarcasm in low-resource languages is challenging because of data scarcity, complexity of language, and cultural aspects. This paper fills the gap in detecting sarcasm automatically from Marathi tweets by using a framework that integrates dataset augmentation, advanced deep learning techniques with Attention Mechanisms. The methods for augmenting data to enhance the small size of the dataset included synonym replacement, back-translation, noise injection, and contextual augmentation, resulting in 13,438 total samples. The models were evaluated using deep learning approaches, with the proposed hybrid CNN-LSTM with attention model achieving the highest performance (Accuracy: 0.9948). Attention Mechanisms allowed the model to focus attention on opposing phrases or words that are exaggerated to indicate sarcasm. Explainable AI (LIME and SHAP) was used to understand the reasons for predictions made by the model and showed that irony, code-mixing, punctuation cues, and positive words are the top contributing features for sarcasm detection. This study shows that a combination of augmentation, advanced deep learning, and explainability can create high-performance interpretable systems for detecting sarcasm in low-resource languages, thus improving the ability to identify sentiment in social media content in Marathi.

Keywords: Marathi NLP, Dataset Augmentation, Deep Learning, Attention Mechanism, XAI

1. Introduction

As millions of social media users express their feelings via emotional, humorous, exaggerated, sarcastic and indirect messages in an informal format, Sarcasm detection has become an important element of the research agenda in Natural Language Processing. Social Networking Sites such as Twitter and Facebook provide a forum for individuals to express their thoughts and opinions using humorous, ironic, and/or exaggerative language. These networks create difficulties for automated systems designed to accurately detect user sentiment or intent from these postings. Marathi and similar lower-resource languages have unique challenges to building datasets due to the existence of multiple spelling forms and different ways of writing (as in multiple languages mixed together). Additionally, cultural aspects of the language/nation can also complicate things for traditional machine learning techniques. Improving sentiment analysis through improved sarcasm detection will greatly impact the ability to monitor social media effectively, mine opinions, and create better interactions between humans and machines across multi-lingual environments.

1.1. Background

The proliferation of social media has provided individuals with the ability to communicate across language and cultural boundaries and therefore emphasizes the immediate need for improved natural language processing tools to interpret this multi-lingual environment (Kumar et al, 2023). Among many of the effects that this growing diversity produces, one of the most significant is the challenge of identifying sarcasm due to the complex contextual and cultural factors that determine sarcastic remarks. In addition to requiring an ironical use of language to express an opposite actual meaning, sarcasm is particularly challenging to identify in languages with rich cultural diversities (Pandey et al, 2021). In the context of Dravidian languages, the task of identifying sarcasm becomes extremely difficult since Dravidian languages have very little in

common with English. People of southern India mostly speak these languages and have unique linguistic features, structures, and cultural references that are not always straightforwardly interpretable by standard NLP models. As a result, sarcasm expressed in these languages may be misinterpreted by conventional sentiment analysis systems, leading to inaccuracies in understanding and responding to user interactions. (Yue et al, 2024)

The use of sarcasm can significantly change the meaning of what is said. In the areas of sentiment analysis, opinion mining, analysis of political discourse, and monitoring of the internet, sarcasm will often cause confusion for classifiers because it conveys positive related content but has the opposite connotation; similarly sarcastic statements that are interpreted to be negative will utilize positive vocabulary. (Maynard & Greenwood, 2014)

1.2. Challenges in Marathi Sarcasm

1. Complex Morphological Constructions

The use of inflectional, derivational and compounding morphologies in Marathi is extensive. A root morpheme will generate a multitude of surface forms, depending on the tense, gender, number and case of the noun. The morphological variations associated with these surface forms often obscure the true meaning of particular statements or thoughts, making it hard for computerised systems to identify sarcasm in particular, where implicit grammatical signals may point to the speaker's intent, rather than direct wording of the text.

2. Usage of Non-Standard Spellings

Informal writing styles are used by many social media users, including phonetic spellings, abbreviations, elongated words, and incomplete/remove transliterations (e.g. Romanized Marathi). Non-standard spellings create data sparsity, which makes tokenization and feature extraction ineffective, thereby lowering the accuracy of sarcasm detection systems.

3. Incidental Use of Emojis

Emojis are often used with text to show emotion, irony or as an emphasis. In Marathi social media, emojis can be used to support the sentiment of a text or deliberately contradict a text for a sarcastic effect. Due to the variability and context dependency of emoji use, it can be difficult to develop models that can accurately identify the meaning of an emoji (i.e., whether the emoji is meant to support or contradict the person's sincerity).

4. Code Mixing in the Language

The majority of Marathi social media messages contain a blend of Marathi and/or English or Hindi in the same sentence or discourse. In fact, code-mixing makes it difficult to identify language and perform sentiment analysis because the grammatical structure and semantic coherence have been disrupted by the use of more than one language in the same way. In addition to these problems, sarcasm expressed in vernacular (mixed-language) constructions is difficult to detect, due to the differing norms of the languages used and contextual changes.

5. Culturally Based Humor

Cultural knowledge, social conventions, expressions unique to specific areas, are the main references in Marathi sarcasm due to the major amount of implicit cultural context needed to understand it as well. Because Computational Models do not typically possess cultural and pragmatic comprehension, they will struggle to recognise the various subtle forms of sarcasm contained within the Marathi language.

2. Literature Review

(Bagate et al, 2025) examined that sentiment analysis (also known as opinion mining) is intended to measure an individual's feelings about a specific topic (e.g., a movie, political party, etc.). The use of sarcasm can create a sense of irony or contempt by providing an account that conveys an attitude of superiority or scorn through the use of insulting language. The use of sarcasm alters the polarity of positive sentiments to negative and vice-versa. The goal of this proposed research is to use a domain-specific analysis to create a sarcasm detection model based on the use of an explainable artificial intelligence (XAI) approach that serves as a verification of the results of this classification. Using a method of providing value to domain-specific text through the use of an explicable approach. In terms of methodology, the study conducted on political

content found on Reddit provides a foundation for the study of sarcastic text available in social media. Natural language processing (NLP) plays a crucial role in sentiment analysis and sarcasm identification. The proposed methodology involves detecting sarcasm through a weighted average approach utilizing long short-term memory (LSTM) and support vector classifier (SVC). The built system employs a novel technique to detect sarcastic terms within sentences or predict sarcastic phrases. The proposed strategy employed an Explainable Artificial Intelligence (XAI) methodology to detect sarcasm inside a particular political text. Counterfactual explanation, as a component of XAI, was utilized to discern sarcastic phrases in the provided text, which were assigned greater significance by the model during training. The system achieved a weighted average F1 score of 75.75%.

(Pakray et al,2025) examined that the Natural Language Processing (NLP) was very well-established, the majority of its advancement has largely benefited only the high-resource languages that had a wealth of annotated data available for training purposes. Unfortunately, for low-resource languages, due to the scarcity of actual large-scale annotated datasets and generally available tools, it makes the implementation of data-driven models very difficult to implement. The complexity of their grammar, diversity of their vocabulary, and the unique social situations in which they exist only complicate the application of standard NLP techniques. Recently, solutions had been developed to try to address these issues, including the development of datasets for low-resource languages based on community-driven collection methods, as well as transfer learning from existing multilingual models such as mBERT and XLM-R. Additionally, there were now multimodal methods available that had been shown to improve performance on various applications, including Sentiment Analysis, POS tagging, Hate Speech Detection, Machine Translation, etc. The continued development of NLP technologies that support low-resource language groups was critical to achieving digital equity and preserving language diversity.

(Dhumpati et al, 2025) suggested that the daily use of social media was now a normal part of everyday life. Many people use social media to share their opinions with others and gather information and news. Nonetheless, social media creates an environment for cyberbullying, which was defined as insulting or harassing people via electronic means, and contains many examples of this type of activity. Identifying sarcasm in these texts can be difficult and may necessitate the use of a sophisticated automated detection system. To help address this problem, we have created a deep learning-based model that uses a convolutional neural network (CNN) to extract features and an attention-based bi-directional long short-term memory (BLSTM) combined with gated recurrent unit (GRU) model to classify these posts as sarcastic. The datasets we created were created from two different sources: public data from Kaggle, as well as news data from multiple sources. Additionally, we included additional natural language processing (NLP) features in our datasets to enhance model performance. An evolutionary strategy to reduce diversity (ESRD) optimizer was used to improve the performance of our models over and above existing state-of-the-art models for detecting both sarcasm and instances of online cyberbullying.

(Golbazi, 2025) discussed a hybrid ACSA model built using McDonald's tweets (as well as other sources of data) with a dataset of 200,000 unique tweets. The hybrid ACSA model consists of BERT, BiLSTM and the Transformer architectures and uses SMOTE algorithm to improve the classification performance. The thesis compares the performance of the hybrid ACSA model with both BERT and BERT + BiLSTM and demonstrates that it produced the highest F1-scores calculated for aspect category (0.78), thematic category (0.85), and sentiment classification (0.86). LIME, SHAP and Attention Visualization techniques were used in conjunction with the thesis to improve the ability for explanation of the model. An analysis of the collection of 200,000 tweets from McDonald's found sentiment trending from negative to neutral and persistent criticism related to CSR, Food, Products, and Customer Service was validated by evaluating the tweet content using the BERTopic algorithm.

(Chauhan & Kumar, 2024) suggested that the rise of social media platforms such as Facebook, Instagram, YouTube, and X had made the world a smaller place by allowing us to communicate with different people around the globe who speak a variety of languages, it had also created challenges related to detecting sarcasm. Sarcasm typically uses an ironic tone that could not be understood without noting the cultural backgrounds of each language being spoken; therefore, this was somewhat difficult when working within a multilingual environment. Although language detection could easily be researched in English-speaking countries, to date there had been limited

research in Dravidian languages, such as Tamil and Malayalam, due to their unique morphology and code-mixing behaviours. In this paper, we review how we had used Transformer-based models (Distil-mBERT, mBERT, and RoBERTa) to aid in detecting sarcastic comments, and how we had positive results and had been encouraged to continue researching multilingual sarcasm using Transformer models.

(Sree et al, 2024) examined that the advent of social media connectivity had created an environment in which individuals could seek validation from a global audience, in contrast to their previous limitations to local communities. In the contemporary landscape dominated by social media, where individuals seek validation and counsel from others, manual analysis of extensive conversations was impractical. To tackle the difficulty of sentiment analysis, we propose utilizing Linear SVC, Logistic Regression, Gradient Boosting Classifier, AdaBoost Classifier, and HistGradient Boosting Classifier in conjunction with lexical analysis. The HistGradient Boosting Classifier was the optimal model for sentiment analysis, achieving the highest mean accuracy of 0.935700 and a ROC AUC mean of 0.905685, signifying exceptional performance in classification accuracy and class differentiation. This method was efficient for analyzing live social activity material, facilitating the understanding of individuals' emotions and their involvement with the content, hence enabling content optimization and the refinement of audience engagement strategies.

(Patil & Kolhe, 2023) examined that the Natural language processing (NLP) had been challenged to determine the real meaning of statements containing sarcasm, making sarcasm detection an area of growing research interest. While there were many existing studies on sarcasm detection using English and several other languages, very few studies had been conducted in the area of sarcasm identification in Indian Languages. The Marathi Language was the third most widely used language in India, yet there had been very little research that explores how to detect sarcasm in Marathi. The majority of the existing methodologies focus on the textual cues, and neglect the additional information that was provided through the use of emojis. Emojis often provide emotional and/or contextual indicators, so they could be very useful tools for detecting sarcasm. The research presented here investigates how to use emojis to improve sarcasm detection in Marathi, and demonstrates that emoji features can improve the performance of several machine learning algorithms for sarcasm detection.

(Tan et al, 2023) examined that the Twitter and Facebook serve as primary platforms for posting emotional responses to daily events, making sentiment analysis a growing field of interest for marketing, behavior monitoring, and new ways to manage public health. At the same time, many people express sarcasm online, where an apparently positive comment may carry a distinctly negative meaning (and thus create challenges for traditional classifiers). Deep learning approaches had resulted in significant advancements; however, existing models continue to perform poorly when trying to determine whether sarcastic negative sentiment was present. Thus, the objective of this paper was to develop a deep neural network-based multi-task learning framework to address both tasks as closely related. The proposed framework produced a performance improvement of three percent and achieved an F1 score of 94 percent.

Table 1: Literature Review

Author (Year)	Key Techniques	Outcomes	Limitations
Bagate et al. (2025)	LSTM + SVC ensemble; XAI	F1-score: 75.75% for political sarcasm detection; Provides explainable predictions.	Performance is moderate (75.75%); Domain-specific (political Reddit data).
Pakray et al. (2025)	Review of transfer learning (mBERT, XLM-R); Community datasets	Highlights solutions for low-resource languages.	Review paper – no new model or empirical results reported.

Dhumpati et al. (2025)	CNN + Attention-based BiLSTM-GRU; ESRD Optimizer	Outperforms state-of-the-art in sarcasm/cyberbullying detection.	Model complexity is high; No specific performance metrics or language focus provided.
Golbazi (2025)	Hybrid BERT+BiLSTM+Transformer (BBT); SMOTE; XAI (LIME, SHAP)	Strong F1-scores (0.78-0.86) for aspect-based sentiment analysis.	Focused on English (McDonald's tweets); Not directly applicable to low-resource languages.
Chauhan & Kumar (2024)	Transformer models (Distil-mBERT, mBERT, RoBERTa)	Positive results for multilingual sarcasm detection in Dravidian languages.	Lacks specific quantitative results; General findings without detailed analysis.
Sree et al. (2024)	Ensemble classifiers (HistGradient Boosting, etc.) + Lexical features	High accuracy (93.57%) and ROC AUC (0.9057) for sentiment analysis.	Focuses on general sentiment, not sarcasm specifically.
Patil & Kolhe (2023)	Traditional ML models enhanced with emoji features	Demonstrated that emoji features improve sarcasm detection performance in Marathi.	Uses traditional ML (not deep learning); No specific performance metrics reported.
Tan et al. (2023)	Deep neural network multi-task learning framework	High F1-score (94%) for sarcasm & sentiment; 3% performance improvement.	Focused on English; High performance may not generalize to low-resource languages.

3. Problem Statement

Sarcasm detection has become significantly more sophisticated for high-resource languages with the development of advanced deep learning techniques and the development of Explainable Artificial Intelligence (XAI). However, there is a lack of development for sarcasm detection in the Marathi language (a major Indian language with very few resources available). The main reason for this gap in development is the fact that there is currently no good-quality, interpretable and high-performance system available to detect sarcasm for Marathi tweets. This problem is exacerbated by the fact that the scarcity of annotated datasets prevents the development of any form of a machine learning-based sarcasm detection system. In addition, the informal nature of Marathi social media text presents a unique challenge to the identification of sarcasm, as it is characterised by code mixing, non-standard spellings and culturally dependent linguistic nuances. Previous research has also relied on a traditional approach to developing machine learning algorithms for this purpose, thus further hindering any progress in this area. The current research aims to address the issues described above and proposes an integrated framework in order to improve both the performance of sarcasm detection and the interpretability of the detection system. This framework combines dataset augmentation (to address the data limitations), advanced deep learning techniques (which use attention mechanisms to model contextual and cultural cues) and explainable AI methods in order to improve the performance and also to allow the user to understand the rationale behind the predictions made by the sarcasm detection system.

4. Research Methodology

This study used a research method that will produce an effective, interpretable system for detecting sarcasm in Marathi tweets. The entire analysis is conducted using NLP processing solely focused on text and augmenting the dataset through deep learning employing an attention mechanism and explainable AI. The dataset is composed solely of text, and it doesn't include images, audio, or any other types of data.

4.1. Dataset Description

In this paper, we provide a comprehensive process for detecting text-based sarcasm on Twitter in the Marathi language using three major steps: creating a larger dataset through augmentation of existing datasets (creating more examples), preprocessing the text with NLP techniques, and using deep learning methods. All steps of our proposed method are identical to the dataset creation and processing steps used in this experiment.

Class Distribution (Before Augmentation)

Prior to data augmentation, the dataset was severely skewed; 5880 non-sarcastic items versus 2520 sarcastic items. This indicated that the machine learning algorithms would be more inclined to classify an item as non-sarcastic because of the disproportionate amount of samples available to them for training. The classification algorithm's capability of detecting sarcasm was thereby negatively impacted. Therefore, it was essential to perform augmentation on the dataset so that the classes could be balanced and the model would be more robust and equitable in providing classifications.

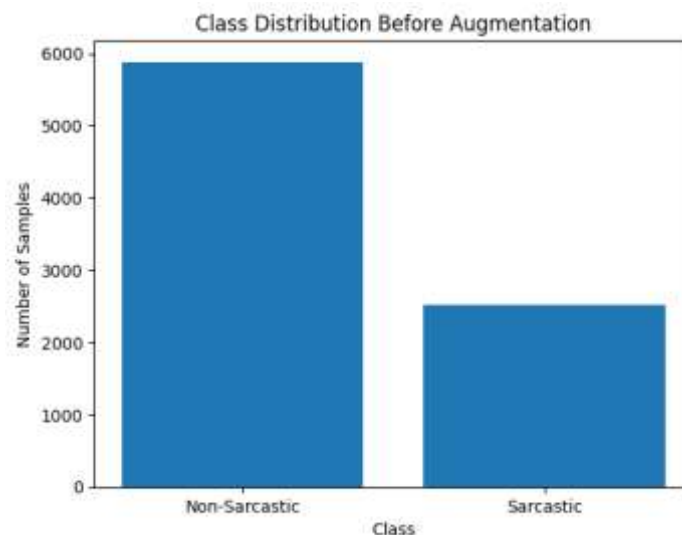


Figure 1: Class Distribution before augmentation

4.2. Data Preprocessing

A. Cleaning and Normalizing

The original text from the social media posts had to go through a process of cleaning to remove any unnecessary material (noise) or irrelevant data. This included removing URLs, user mentions, and any repeated characters because they did not add to the overall sarcastic intention of the post; they also have a negative effect on the ability of the learning algorithms to correctly identify sarcasm. The researcher also excluded emojis from this research because it was set up as a text-only analysis. By doing so, this removes any opportunity for the sarcastic intent of the post to be influenced by the presence of an emoji. All informal spellings commonly used in electronic communication (e.g., u, ur) were converted to their proper dictionary spelling forms. This was done to keep the words in the same form throughout the data set.

B. Tokenization

Subword tokenization is accomplished using techniques such as WordPiece or Byte Pair Encoding (BPE). These approaches are particularly useful in handling the unique nature of Marathi social media text, where there is considerable code-mixing and many non-standard word forms. Additionally, they mitigate the out of vocabulary problem by dividing linguistically relevant items into smaller parts and capturing them across different languages in one sentence.

C. Label Encoding

The task of detecting sarcasm was framed as a binary classification problem, where sarcasm was assigned the label of 1 and non-sarcasm was assigned the label of 0. The ability to encode these sarcasm instances numerically so that they could be processed efficiently by machine-learning and deep-learning algorithms also provided clear class distinctions between the two classes.

D. Feature Engineering

A number of handcrafted features were created to improve the model's detection of sarcasm. For example, punctuation count showed that sarcasm was associated with overly exaggerated forms of expression. In addition, the language mixing ratio indicated how much code-mixing occurred within a single sentence, which is typical for Marathi social media messages. Indicators of contrast were created to indicate instances where the meaning or feeling of a word or phrase had changed to the opposite of what was intended by the author; these often indicate sarcasm being used. The token length was also included to account for differences in sentence structure and levels of complexity or elaboration when expressing thoughts or ideas.

4.3. Dataset Augmentation

To address the dataset imbalance and improve the model's ability to generalize, four augmentation techniques were used.

Table 2: Augmentation Techniques

Technique	Description
Synonym Replacement	Marathi WordNet synonyms
Back-Translation	Marathi → Hindi → Marathi
Noise Injection	Spelling variants, punctuation
Contextual Augmentation	Minor lexical paraphrasing

Through using data augmentation methods, the total number of samples in the dataset increased to 13,438 total samples consisting of 7,558 sarcastic samples and 5,880 non-sarcastic samples. This method of augmentation was utilized to address the original class imbalance and still retain the natural distribution of data as opposed to imposing an artificial 50-50 split between the two classes (sarcastic vs non-sarcastic). With a controlled increase in the number of sarcastic examples, this increased variability and diversity within the augmented dataset allows for the models to learn more robust sarcasm-based patterns. Therefore, having this balanced but authentic distribution enhances the models' ability to generalize while also reducing the potential bias associated with over-synthetic or highly uniform proportion sizes within the original dataset.

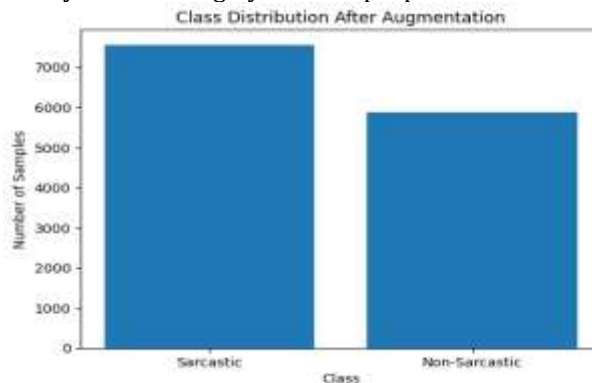


Figure 2: Class Distribution after augmentation**4.4. Text Vectorization and Embedding**

By employing a deep learning methodology, dense word embedding has been generated based on the manifestation of both contextual and semantic relationships among the various tokens, with an embedding matrix created based on the padded token sequences used for training and to develop distinct, distributed representations of Marathi and code-mixed social media postings. The number of tokens allowed in the maximum sequence length was set at 50 to provide standardization across the various models used and better facilitate batch processing. As such, the embedding-based representations have provided those deep neural networks with an ability to identify and represent nuanced and subtle semantic distinctions, as well as contextual cues, that are crucial for detecting instances of sarcasm.

5. Experimental Setup**5.1 Dataset Description and Augmentation**

To perform an experimental evaluation using a curated corpus containing a total of 8400 Marathi Tweets obtained from different social media channels. In order to remedy the problem of inherent class imbalance and increase the generalization ability of the models, the data was expanded from 8400 Tweets to 13438 Tweets. The total number of augmented Tweets includes 7558 instances of sarcasm and 5880 instances without sarcasm, so the entire dataset has an accurate representation of the class distribution as opposed to an artificially created balance.

In addition, after pre-processing and normalisation of the data, there was a controlled vocabulary of 4195 tokens developed. This controlled vocabulary also allowed for the consistency of all model inputs, since a maximum input of 50 tokens per Tweet was maintained in all experiments. By employing the use of "back-translated" and "synonym-replacement" augmentation techniques as a means to add diversity to the language used in the Tweets, back-translation produced the greatest improvement of representations for both sarcasm and the ability to reduce potential overfitting.

5.2 Attention Mechanism and Transformer Encoding

To improve the model's ability to focus on important linguistic elements within an input text, the authors used an Attention Mechanism that uses a Query (Q), Key (K), and Value (V) representation of the input text based on Hidden State representations. The attention weights were calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

With this formulation, the model emphasizes contrastive phrases, intensifiers, and other exaggerated language markers associated with sarcasm. For Transformer-based representations of tweets, the model used a sequence of token embeddings in the form of $\{X_1, X_2, \dots, X_n\}$. The Transformer encoder performed this multichannel attention via the multi-head attention mechanism and the feed-forward layers that are based on the transformer encoder.

$$h^{(l)} = \text{MultiHeadAttn}(h^{(l-1)}) + h^{(l-1)}$$

$$z^{(l)} = \text{FFN}(h^{(l)}) + h^{(l)}$$

To stabilize training, residual connections were used. The final pooled representation was then sent to a softmax classifier, which was used for binary classification.

5.3 Loss Function

The binary cross-entropy loss function, which is suitable for binary classification problems, was used to optimize all models. The loss is defined as follows:

$$L = \sum_{i=1}^N \sum_{c \in \{0,1\}} y_{i,c} \log \hat{y}_{i,c}$$

$y_{i,c}$ is the true label associated with sarcasm or a lack thereof (non-sarcasm) whereas y is the estimation of probability for a specific class. This loss definition is designed to reward good probability estimates and help ensure consistent training convergence.

5.4 Integration of Explainable AI

The purpose of integrating explainable AI (XAI) into the experimental design was to increase the interpretability and transparency of the proposed framework. The individual tweet-level predictions from the classifier were assessed using Local Interpretable Model-Agnostic Explanations (LIME). LIME enabled us to determine which lexical components were most impactful in terms of the prediction made by the model and thus how each tweet was classified. SHapley Additive exPlanations (SHAP) was used to calculate the global importance of features across the entire dataset. The results from SHAP analysis provided evidence that sarcasm detection uses certain recurring indicators in the language used to express sarcasm; specifically, the presence of ironic positive language, regional informal/slang language, punctuation emphasis, and contrast regarding the information being referenced.

With respect to attention-based neural network models, heatmaps (or attention maps) were produced to visualize how much attention was directed at specific tokens for each input tweet. The attention heatmaps showed that the models directed a substantial portion of their attention toward the contrastive monochromatic phrases, exaggerated adjectives, and especially emphasis patterns in the tweets; all of these types of tokens provide linguistic cues that are relevant for sarcasm detection.

6. Results & Discussion

6.1 Dataset Statistics and Summary

The information presented in this piece gives general statistics of the data being tested through the experimental evaluation. The source file contained 8,400 unique Marathi-language tweets. However, to fix any issues with class imbalance and generalization, the number of documents was increased to 13,438 by adding additional samples of each class. All data and descriptive statistics for the entire dataset can be found in Table 3.

Table 3. Dataset Statistics and Summary

Metric	Value
Original Samples	8,400
Samples After Augmentation	13,438
Sarcastic Samples	7,558
Non-Sarcastic Samples	5,880

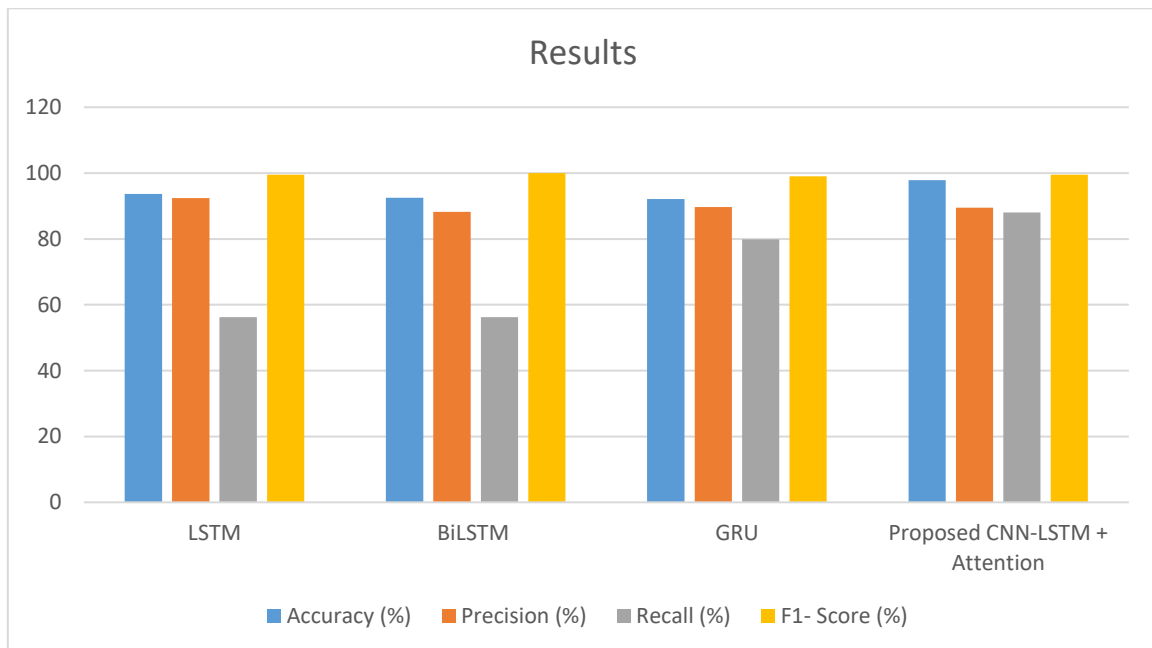
The data showed that augmentation significantly increased the number of sarcastic examples, while also maintaining a realistic distribution of classes.

6.2 Comparative Analysis of Deep Learning Models

As illustrated in the table, the performance of the deep learning architectures (as measured by accuracy, precision, recall, and F1-score) was evaluated, providing a complete understanding of how well each architecture performed at classifying records.

Table 4. Deep Learning Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)
LSTM	93.70	92.55	92.14	93.85
BiLSTM	92.40	88.28	89.67	89.47
GRU	56.25	56.25	100.00	72.00
Proposed CNN-LSTM + Attention	99.48	100.00	99.07	99.53

**Figure 4 Performance Analysis**

Based on the comparison results, the CNN-LSTM with Attention model demonstrates superior performance compared to all other architectures with respect to the evaluation metrics. Specifically, it has the highest level of classification accuracy (99.48 %) and F1-Score (99.53 %) across all evaluation metrics, indicating it has a very good ability to classify instances into the correct category while maintaining a good balance of precision and recall. The model's perfect level of precision (100.00 %) indicates that it makes virtually no false positive errors and its high level of recall (99.07 %) indicates that it is highly robust in detecting instances of sarcasm. In comparison, the standard LSTM model has shown competitive performance, but at a much lower level than the Attention model, indicating that LSTMs are limited in their capabilities to identify and learn complex contextual cues and contrastive cues. The BiLSTM model is even less effective than the LSTM and is particularly weak in terms of precision and recall because it does not use a mechanism to explicitly focus attention on specific parts of the input sequence. On the other hand, although the GRU model achieves perfect recall, its extremely low levels of precision and accuracy indicate that it is heavily biased toward the sarcastic class, which shows that the GRU architecture is not effective for use as a reliable sarcasm detector in this specific scenario.

6.3 Impact of Dataset Augmentation

- **Improved Model Generalization:**

By adding synthetic diversity to the dataset, the models were better able to generalize to new and varied linguistic patterns often found in social media text.

- **Enhanced Representation of Sarcastic Class:**

The augmentation process increased both the number and variety of sarcastic examples, which were underrepresented in the original dataset. This, in turn, reduced the bias toward predictions that didn't include sarcasm.

- **Reduction in Overfitting:**

The expanded dataset exposed deep learning models to a broader range of sentence structures and expressions. This led to reduced overfitting and more stable training.

- **Improved Recall of Sarcastic Samples:**

A notable improvement was observed in the recall metric, indicating that the models became more effective at correctly identifying sarcastic content after augmentation.

- **Effectiveness of Augmentation Techniques:**

Among the applied strategies, back-translation and synonym replacement contributed most significantly to increasing linguistic diversity and improving overall classification robustness.

6.4 Findings from Explainable AI Analysis

Explainable AI techniques were employed to validate whether the models learned linguistically meaningful patterns rather than spurious correlations.

A. LIME Analysis:

Local explanations generated by LIME revealed that model predictions were strongly influenced by:

- Contrastive terms
- Sarcasm-indicative intensifiers
- Mixed-language triggers

These features align closely with known linguistic markers of sarcasm in Marathi social media discourse.

B. SHAP Analysis:

Global feature importance analysis using SHAP identified the following dominant contributors:

- Ironic positive words used in negative contexts
- Likeable regional slang expressions
- Punctuation emphasis
- Contrastive linguistic constructs

These findings confirm consistent feature relevance across the dataset.

7. Conclusion & Future Scope

In this research, an integrated framework for high-performance and interpretable sarcasm detection in Marathi tweets was developed. The use of data augmentation, Deep Learning and Explainable AI (XAI) facilitated the overcoming of both the issues with data scarcity and the transparency of the models' decisions. The linguistically meaningful cues that were of greatest importance (contrastive terms, code-mixing, etc.) were confirmed by the model's outputs, which were validated through LIME and SHAP.

In the future, it is possible that the findings from this research will improve the performance of sarcasm detection through the use of additional multimodal sources of information (such as emojis and images) that provide a greater understanding of the sources of non-textual cues used in social media communications to enhance sarcastic intent. Additionally, future research may experiment with the use of more advanced pre-trained language models to gain a more robust understanding of contextual and semantic relationships in human language, particularly with regard to low-resource languages (such as Marathi). Furthermore, future work will likely apply cross-lingual adaptation techniques to the proposed framework for related low-resource languages, allowing the transfer of knowledge across languages and improving the generalisability of the framework to other languages. Furthermore, deploying the sarcasm detection model in real time and incorporating continuous learning mechanisms to improve the model's robustness to variations in the data collected through user feedback would improve the model's practical and long-lasting applicability.

References

1. Kumar, A., Singh, J. P., & Singh, A. K. (2023). Explainable BERT-LSTM stacking for sentiment analysis of covid-19 vaccination. *IEEE Transactions on Computational Social Systems*.
2. Kumar, A., & Singh, J. P. (2022). Deep neural networks for location reference identification from bilingual disaster-related tweets. *IEEE Transactions on Computational Social Systems*, 11(1), 880-891.
3. Kumar, A., Singh, J. P., Rana, N. P., & Dwivedi, Y. K. (2023). Multi-channel convolutional neural network for the identification of eyewitness tweets of disaster. *Information Systems Frontiers*, 25(4), 1589-1604.
4. Pandey, R., Kumar, A., Singh, J. P., & Tripathi, S. (2021). Hybrid attention-based long short-term memory network for sarcasm identification. *Applied Soft Computing*, 106, 107348.
5. Pandey, R., Kumar, A., Singh, J. P., & Tripathi, S. (2025). A hybrid convolutional neural network for sarcasm detection from multilingual social media posts. *Multimedia Tools and Applications*, 84(16), 15867-15895.
6. Yue, T., Shi, X., Mao, R., Hu, Z., & Cambria, E. (2024, May). SarcNet: a multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 14325-14335).
7. Maynard, D. G., & Greenwood, M. A. (2014, March). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.
8. Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014, October). "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 116-126).
9. Kim, Y., Yim, J., Yun, J., & Kim, J. (2019). Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 101-110).
10. Barbieri, F., Ballesteros, M., & Saggion, H. (2017). Are emojis predictable?. *arXiv preprint arXiv:1702.07285*.
11. Barman, U., Das, A., Wagner, J., & Foster, J. (2014, October). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 13-23).
12. Attardo, S. (2024). *Linguistic theories of humor* (Vol. 1). Walter de Gruyter GmbH & Co KG.
13. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. 2016.
14. Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1), 3.
15. Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118(4), 374-386.
16. Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of pragmatics*, 32(12), 1777-1806.
17. Gibbs, R. W., & Colston, H. L. (Eds.). (2007). *Irony in language and thought: A cognitive science reader*. Psychology Press.
18. Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5), 1-22.
19. Davidov, D., Tsur, O., & Rappoport, A. (2010, July). Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107-116).
20. Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74, 1-12.
21. Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 704-714).
22. Ptáček, T., Habernal, I., & Hong, J. (2014, August). Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 213-223).

23. Bagate, R. A., Joshi, A. S., Kadam, A., Choubey, C. K., Sable, N., Kumar, A., ... & Nandan, D. (2025). Sarcasm Detection an Explainable AI Approach for Reddit Political Text. *Mathematical Modelling of Engineering Problems*, 12(1).
24. Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2), 183-197.
25. Dhumpati, R., Sasi, A., Basha, S. J., Vadhri, S., Saroja, P., Alluri, S., & Vatambeti, R. (2025). Enhancing sarcasm detection in sentiment analysis for cyberspace safety using advanced deep learning techniques. *Scientific Reports*, 15(1), 22681.
26. Golbazi, J. (2025). Development and Application of a Hybrid Deep Learning Model for Aspect-Based Sentiment Analysis (ABSA): A Comparative Study of McDonald's Tweets Across Two Time Periods.
27. Chauhan, S., & Kumar, A. (2024). A Transformer-Based Model for Detecting Multilingual Sarcasm in Social Media Posts. In *Forum of Information Retrieval and Evaluation FIRE-2024*.
28. Sree, M. R., Siddhartha, M., Reddy, P. V. V., & Belwal, M. (2024, June). Sentiment Analysis and Likes Prediction from Social Media Comments: An ML approach. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-8). IEEE.
29. Patil, P. K., & Kolhe, S. R. (2023, June). Sarcasm Detection for Marathi and the role of emoticons. In *International Conference on Data Intelligence and Cognitive Informatics* (pp. 193-204). Singapore: Springer Nature Singapore.
30. Tan, Y. Y., Chow, C. O., Kanesan, J., Chuah, J. H., & Lim, Y. (2023). Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless personal communications*, 129(3), 2213-2237.