VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 2, April-June 2026 pp: 01-16

Model Interpretability with XAI for Sarcastic Behavior Detection in Low-Resource Roman Urdu Language using Machine Learning and Ensembled Approaches

Amirita Dewani [1*], **Memoona Sami** [1], **Dua Agha** [2], **Veena Kumari** [1], **Mariam Memon** [1]

[1]Department of Software Engineering at Mehran University of Engineering and Technology, Jamshoro, Pakistan;
[2]Computer Science & Information Technology Department, NED University of Engineering and Technology, Karachi, Pakistan

Keywords: *Sarcasm detection, Asian Regional Languages, Roman Urdu, Low-resource languages, Artificial Intelligence, Ensemble learning, Explainable AI (XAI).*

Journal Info:
Submitted:
February 01, 2026
Accepted:
April 05, 2026
Published:
April 10, 2026

Abstract This work proposes a framework for sarcasm detection in Roman Urdu using machine learning and ensemble approaches. For dataset preparation, we extracted data from X and manually annotated a corpus of 11,320 tweets into two classes, i.e., sarcastic and non-sarcastic. Various classifiers were implemented using lexical features related to sarcastic patterns. The proposed preprocessing framework and ensemble-based model use language-related normalization strategies such as spelling variation standardization, and slang expansion to address orthographic variation and code-mixing within the text for sarcasm detection. Experimental findings show that ensemble learning models, especially Random Forest and XGBoost obtained the most accurate results with an accuracy of 87% and an F1-score of 88%, thus providing a reliable baseline on sarcasm detection using South Asian Roman Urdu language. These findings emphasize the potential of ensemble approaches to process and handle the complexity of sarcasm in low-resource code-mixed native languages. Furthermore, in order to enhance transparency and trust in the model's decisions, we employed the Explainable AI (XAI) technique, particularly LIME (Local Interpretable Model-agnostic Explanations), to interpret the predictions of our best-performing model. These findings emphasize the potential of ensemble approaches to process and handle the complexity of sarcasm in low-resource code-mixed languages, and XAI offers essential insights into the process of the model's decision-making.

***Correspondence author email address:** amirita@faculty.muet.edu.pk
DOI: 10.21015/vtse.v14i2.2343

1 Introduction

The exponential growth of social media platforms such as X (formerly Twitter), Facebook, and Instagram has transformed the way individuals share opinions, reviews, and commentary on diverse topics and various issues such as politics, entertainment, and consumer products. Such platforms have a tremendous volume of user-generated content that offers an invaluable resource for mining public sentiment and behavioral patterns. In this context, Sentiment Analysis (SA) has emerged as a key area in Natural Language Processing (NLP) research, which is defined as the computational study of opinions, sentiments, and emotions expressed in text [1, 2]. However, sentiment analysis becomes significantly complicated and more challenging when

confronted with sarcasm, a form of figurative language where the intended meaning contrasts with the literal interpretation [3]. Sarcasm can invert the polarity of an expression from a positive statement to a negative or vice versa, thus leading to misclassification in traditional sentiment models [4, 5]. In recent years, sarcasm detection has attracted increasing research attention because of its impact on improving sentiment analysis, opinion mining, and human-computer interaction [6, 7]. Although significant progress has been achieved on high-resource languages like English [8, 9], relatively, the work on low-resource languages remains limited due to a scarcity of annotated data in languages beyond English, the lack of related tools and resources, and unique challenges posed by complex morphology and informal syntax [10]. Specifically, Romanized script versions of Urdu pose multiple challenges for computational processing, such as:

1. Non-standard orthography, where the arrangement of letters in the same word may vary in different ways or may have multiple spellings (e.g., “acha” or “achha”).
2. Code-mixing with English that contributes to syntactic and semantic complexity.
3. Informal constructs that carry contextual meaning, such as abbreviations, emojis, and hashtags.
4. Rich morphological variation typical in Urdu language structure [11].

Roman Urdu has gained increasing traction on social media platforms, yet it remains underrepresented in sarcasm detection research. The majority of the existing works dealing with low-resource sarcasm detection have considered standard Urdu or other morphologically rich languages, such as Arabic, Hindi, and Pashto [7, 10, 12]. In addition, a significant portion of the current research is based on single-model classifiers, which can be inadequate to identify the multidimensionality of the linguistic and contextual cues present in sarcastic utterances. Recent findings suggest that ensemble learning approaches integrate the strengths of multiple classifiers to outperform individual models by offering complementary decision boundaries in complex linguistic settings [13, 14]. This research addresses these gaps by developing a manually annotated Roman Urdu sarcasm dataset by scraping X, implementing machine learning methods and ensemble approaches with Roman Urdu-specific preprocessing, along with assessing and evaluating their performance using various lexical features. By systematically comparing multiple algorithms and ensemble configurations, this study aims to establish a robust performance benchmark for sarcasm detection in Roman Urdu that contributes to broadening the scope and advancement of NLP for low-resource languages. In summary, this study makes the following contributions:

1. Dataset Development: Construction of a manually annotated Roman Urdu sarcasm dataset (extracted from X) that will address an imperative lack of research resources on low-resource languages. The dataset comprises 11,320 comments annotated manually for this work.
2. Language-Specific Preprocessing: Implementation of preprocessing methods tailored to Roman Urdu to address and deal with orthographic variation, code-mixing, and informal textual patterns.
3. Comparative Evaluation: Systematic comparison of multiple machine learning techniques and ensembled strategies using a combination of lexical features.
4. Performance Baseline for a newly adopted Asian language: Setting a solid baseline for sarcasm detection in Roman Urdu, which allows for future directions in multilingual and code-mixed NLP research, typically for regional languages.
5. Model Interpretability: Utilized Explainable AI (XAI) using LIME implementation to provide transparent, human-readable explanations for the sarcasm prediction to foster trust and gain deeper insights into the linguistic cues learned by the model.

Further sections of this paper are organized as follows: Section 2 provides a comprehensive review of the existing literature on developed approaches for detecting sarcasm in different languages globally. It highlights the key findings and gaps in existing work contributions, thus setting up the research baseline and significance. Section 3, in continuity with the findings of the previous section (i.e., Literature Review), defines and discusses the problem being addressed in this research. Section 4 exhibits the methodological framework employed in this work for sarcasm detection in Roman Urdu tweets.

It consists of seven phases, which are further detailed in different sub-sections. The next section, Section 5, narrates and puts light on the research results. Section 6 provides insights into the reasons a model produces a given prediction to establish trust, guarantee fairness, and optimize the system, particularly in subtle social tasks like the detection of sarcasm. Finally, Section 7 concludes the contributions made, provides social implications for the current work, and suggests avenues for future research extensions.

2 Literature Review

Detection of sarcasm in languages has gained increasing academic attention in recent years, but to the best of our knowledge, there has been limited work on the topic applied to resource-poor languages, including Roman Urdu. We have grouped the existing work into dataset construction, classical machine learning approaches, deep contextual models, and ensemble and explainability techniques.

Classical Machine Learning Approaches: From a classical machine learning perspective, this gap can be observed in [10], which directly bridges it by studying the application of nine machine learning algorithms Support Vector Machines (SVM), Random Forests, Decision Trees, K-Nearest Neighbor (KNN), Linear Regression, Naive Bayes, and XGBoost to the detection of sarcasm in Urdu text, thus presenting the first systematic assessment of these models in a low-resource setting. As indicated in the results, SVM is the one that works best with an accuracy of 85%. While these results show that classical machine learning methods can be effective in low-resource settings, such approaches mainly rely on surface-level lexical features and may not fully capture the implicit and context-dependent nature of sarcasm. However, even with such limitations, this observation supports the assertion that knowledge-representation machine-learning methods can reliably be applicable to the peculiar linguistic and cultural characteristics specific to resource-limited languages.

Dataset Construction and Low-Resource Challenges: In terms of dataset construction, [4] composes a balanced corpus of 12,910 manually labeled tweets in Urdu and analyzes a variety of deep learning structures, such as CNN, LSTM, GRU, BiLSTM, as well as CNN-LSTM fastText word embeddings. They also propose a hybrid multilingual BERT (mBERT) -BiLSTM- multi-head attention model that performs well in accounting for sequential and contextual relationships, arriving at a 79.51% accuracy and 80.04% F1 score. This shows that the combination of transformer-based contextual embeddings, recurrence, and attention in low-resource scenarios is beneficial. However, such approaches are typically data-intensive, and their effectiveness in highly variable and non-standardized text such as Roman Urdu remains underexplored.

The study of multilingual and code-mixed resources should be supported systematically, as demonstrated in [12], whose article presents the first publicly released English-Hindi code-mixed sarcasm dataset. The dataset was collected via hashtag-based retrieval and manually verified with a language-level token tagging. Reliability tests of these procedures were high with Cohen Kappa of 0.79, meaning that there was high annotation reliability. The Random Forest classifier gave an average F-score of 78.4% when applied to the dataset, and therefore, this work not only provides a benchmark dataset but also establishes a reproducible methodology for corpus creation in low-resource, code-mixed environments. Nevertheless, differences in annotation schemes, domains, and class distributions across such datasets make direct comparisons difficult and highlight the lack of standardized benchmarks.

Contextual and Deep Learning Approaches: As the literature indicates, with the advancement of deep learning and contextual models, transformer-based architectures are found exemplary in terms of their performance when used to identify sarcasm in short, informal social-media language. According to [5], this ability is depicted by adopting BERT-style bidirectional contextual embeddings to X (Twitter), which demonstrates 97% accuracy on a Kaggle dataset and 92% on unique Twitter data, surpassing LSTM, and other traditional machine-learning baselines. These findings highlight the capabilities of BERT to represent contextualized linguistic formulations with the peculiarity of noisy code-mixed tweets. At the same time, Abuzayed and Al-Khalifa in [7] show how the transformer model can be effectively used on the WANLP 2021 shared task on Arabic sarcasm detection through data augmentation,

augmenting MARBERT in addressing extreme imbalance. An enhanced sarcasm recognition of F1-scores (leading to approximately 15% increase) was achieved using this technique, therefore, offering generalizable methodologies to low-resource, morphologically rich languages. These studies demonstrate the strength of contextual representations in sarcasm detection, but they also indicate that such models are often developed and evaluated outside Roman Urdu settings.

Regarding context-aware sarcasm detection, [15] studied the effect of conversation context on sarcasm identification in conversation based on Twitter and Reddit data sets and provided assessments. The included conversational turns in recent studies have allowed improvement in performance in most instances where BERT had the highest score with both BiLSTM and SVM showing the lowest. According to the authors, sarcasm detection in any bi-lingual text can be achieved using context modeling, but the most appropriate context length hugely depends on the domain. While this challenge is important for Roman Urdu as well, most existing low-resource studies, including the present one, operate primarily at the tweet level.

Ensemble Methods: In terms of ensemble approaches, more recent research shows that ensemble-based methods are best applied in heterogeneous language environments. [13] investigated the topic of cross-domain sarcasm detection through the use of TF-IDF features and various machine-learning models and then combined the results of all the pairs of models using Voting and Stacking of the results. The generated ensembles were consistently better than single-model systems, and it was based on the complementary decision boundaries, which were helpful in complex linguistic environments. Simultaneously, [14] used Word2Vec embeddings along with Multi Weld Matrix Estimation to localize semantic similarity and used those representations in classifiers like SVM and Random Forest. Obtaining 90-96% accuracy on hashtag-labeled tweets, their results support the applicability of semantic similarity in sarcasm detection, though the constantly evolving memes, slang, and emojis are also an ongoing dilemma. These findings suggest that ensemble learning is a promising direction for low-resource and noisy text settings such as Roman Urdu.

Some previous studies have laid down initial procedures upon which sarcasm can be detected in the context of social media. [8] reviewed lexical n-grams, patterns of punctuation, and sentiment polarity changes in balanced openly accessible Czech and English tweets to provide insight. Their empirical results indicate that, despite partial transferability of some classes of sarcasm related features, it would be necessary to use linguistic-specific or domain-specific resources when adapting English trained systems for complex linguistics. [9] compiled a high-quality corpus of sarcasm using The Onion and HuffPost headlines and presented a Hybrid Neural Network that uses joint LSTM, CNN, and attention blocks. They conduct their research using English news headlines, but since it focuses on producing rigorous annotation and contextual modeling, their approach can be utilized by low-resource sarcasm detection research.

Cross-Language and Multilingual Developments: Recent work has started to explore transformer-based architectures for sarcasm and sentiment analysis in other low-resource South Asian languages. For Bengali, a transformer-based generative adversarial model has been proposed for sarcasm detection from social-media comments using limited labeled data, and it reports strong performance on confusing sarcastic text [16]. A recent survey on Bengali sentiment analysis further shows that Transformer-based models such as BanglaBERT now achieve state-of-the-art performance across several Bengali sentiment and emotion benchmarks, while still highlighting persistent data scarcity and linguistic challenges in this low-resource setting [17]. Complementing these efforts, the Ben-Sarc corpus introduces a large-scale, manually annotated dataset of Bengali social-media comments for sarcasm detection and provides baseline results with machine learning, deep learning, and transfer learning models [18]. Other studies fine-tune multilingual transformer models on under-resourced languages and show that these models can achieve higher accuracy than traditional machine-learning baselines for sentiment analysis in noisy social-media settings [19]. Shared tasks on Tamil and other Dravidian languages also make use of multilingual BERT and related transformers for sentiment and sarcasm in code-mixed comments, again confirming the effectiveness of transformer-based contextual

representations in low-resource environments [20]. However, these studies also show that data scarcity, linguistic variability, and cross-domain generalization remain unresolved challenges.

Explainability: From an explainability perspective, in low-resource and code-mixed languages, relatively less attention has been given to understanding why sarcasm detection models make specific predictions. Most of the previous work majorly focuses on predictive performance, rather than interpretability. This makes explainability in the context of sarcasm detection in languages like Roman Urdu an opportunity for exploration.

Research Gap and Motivation: Overall these studies demonstrate the gradual progression from traditional machine learning to advanced transformer-based and ensemble approaches. At the same time, it emphasizes the critical role of high-quality datasets and language-specific adaptations for different regions. The existing work contributed by the research community mostly focuses on the lingua franca [21], which comprises languages spoken globally, unlike native or regional languages, for example, English. We recognize that contextual and transformer-based modeling are extremely important directions, but for low-resource, code-mixed languages, especially Roman Urdu which has diverse stylistic expressions, the most immediate gaps are firstly the lack of a dedicated annotated dataset, secondly the absence of strong benchmark baselines, and finally the limited exploration of ensemble-based and interpretable approaches for sarcasm detection. In this context, the present study focuses on developing a Roman Urdu sarcasm dataset and evaluating the performance of machine learning as well as ensemble methods on it with an initial explainability analysis.

3 Problem Statement

Sarcasm detection in textual data is quite a challenging task due to the fact that it frequently relies on subtle contextual and cultural cues that are difficult to model computationally. While significant progress has been achieved in sarcasm detection for high-resource languages, languages that can be characterized as low-resource, like Roman Urdu, have not been explored to the full extent [4, 10]. Detecting sarcasm is especially tricky in Roman Urdu for NLP models due to its non-standardized spelling, its code-mixing with the English language, as well as rich morphological variation. Moreover, the lack of publicly available annotated datasets and language-specific tools also contributes to the obstacle in the development of effective detection systems and mechanisms [22]. Most existing low-resource sarcasm detection methods have been built on conventional machine learning frameworks trained on handcrafted features. While these approaches achieve moderate performance, they may fail to capture the deeper semantic and pragmatic nuances inherent in sarcastic expressions. Recent advances in ensemble learning have proven to be effective in addressing such challenges by combining multiple models to improve accuracy and robustness [13, 23]. However, their application to Roman Urdu sarcasm detection remains unexplored. Therefore, this study aims to bridge this gap by developing a Roman Urdu sarcasm dataset along with utilizing machine learning and ensemble methods with language-specific preprocessing.

4 Research Methodology

This section outlines the methodological framework employed in this study for sarcasm detection in Roman Urdu tweets. It consists of seven primary stages: data extraction, preprocessing, annotation, data splitting, feature selection, model training (using machine learning and ensembles), and evaluation. Figure 1 represents the methodology.

4.1 Data Acquisition

The dataset was constructed by scraping tweets from the Twitter (currently known as X) platform. A total of 11,320 tweets were initially gathered to ensure a diverse mix of sarcastic and non-sarcastic Roman Urdu text. These tweets were used as raw data in the further preprocessing, annotation, and classification steps. After adhering to the terms of the platform and following best practices of reproducibility in social-media NLP, we kept the content of each tweet and stored a list of their respective ids to allow redistribution and rehydration by other researchers. No user profiling was performed beyond what is intrinsic to the tweet text.

It is important to note that data collected from Twitter/X may not be the full representation of the broader

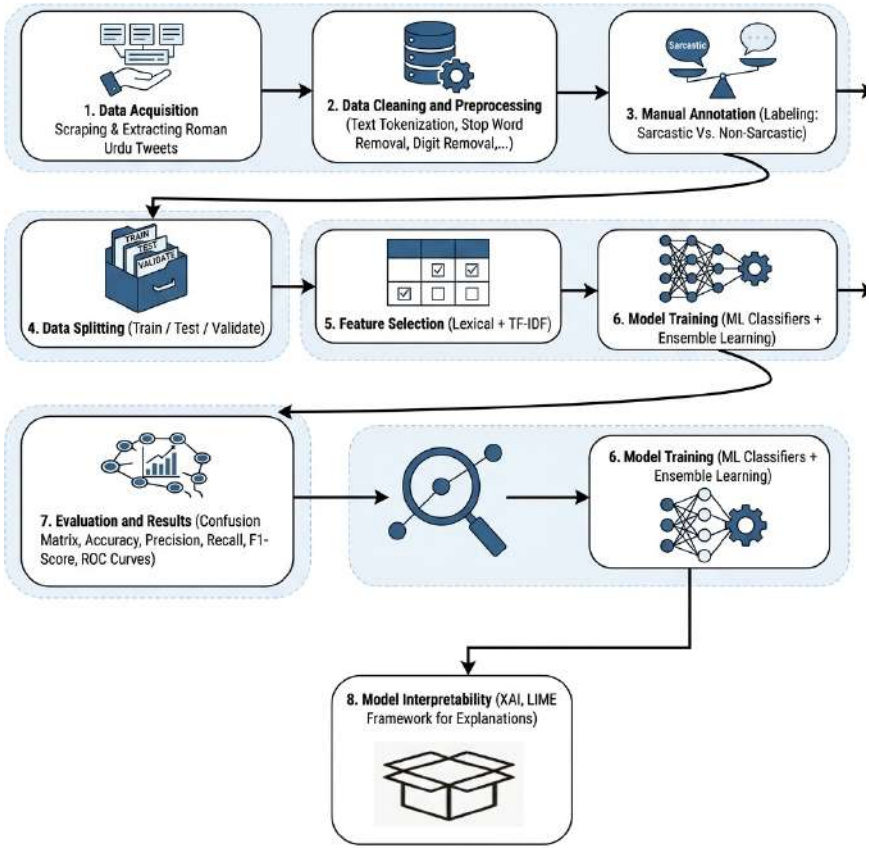


Figure 1. Proposed Methodology for Sarcasm Detection in Roman Urdu Language

Urdu-speaking population because the demographics on the platform are typically dominated by younger, urban, and technologically engaged users. As a result, the language patterns found in this type of material may be distinct from those found in other linguistic domains such as formal text, news discourse, or private messaging platforms. Additionally, social media datasets are affected by time; therefore, data gathered over particular periods or around specific events may overpopulate certain topics, sentiments or ways of expression. The factors may impede researchers from generalizing the findings across settings.

4.2 Data Preprocessing

Preprocessing plays a significant role in textual data since tweets received in their original form may contain noise, e.g., special symbols, links, metadata as well as unconventional spelling, which was, in this case, rectified by pre-processing. Sequentially, the following operations were performed:

4.2.1 Data Cleaning

This phase ensured that only meaningful text was retained in the data. Irrelevant metadata, as well as elements with no use in the text classification, were eliminated. These included user mentions, timestamps, among other structural tags that do not contribute to textual content.

4.2.2 Lowercasing

All characters were converted to lowercase. This eliminated the duplication of information in the data caused by variations in capitalization that tend to be common in informal Roman Urdu written language.

4.2.3 URL Removal

As URLs only point to external web resources and do not provide any lexical or semantic cues related to sarcasm, they were considered noise and were discarded.

4.2.4 Text Tokenization

The cleaned text was then tokenized into smaller units (tokens). Tokenization divided the tweets into words and symbols, which are required in determining the lexical features and classification.

4.2.5 Stop-Word Removal

The elimination of stop words, i.e., high-frequency words that do not impact the sarcastic tone of a sentence, was carried out. Since such words have little to no semantic association with sarcasm detection, they were eliminated from the dataset.

4.2.6 Digit Removal

Numerical digits present in the tweets were removed. Digits were not seen as beneficial to finding out sarcasm since the numbers do not contribute to the linguistic/semantic structure of a sarcastic utterance.

4.3 Data Annotation

Following preprocessing, manual data annotation was performed. Each tweet was assigned one of the two labels:

- 0 - No Sarcastic tweet in Roman Urdu
- 1 - Sarcastic tweet in Roman Urdu

The annotation process involved three annotators, all of whom were native Urdu speakers and graduates. Annotation guidelines were created aimed at providing consistency and were explained to the annotators before they began the task. Every tweet was classified individually by two annotators. In cases of disagreement, the third annotator assessed the case and arbitrated the conflict. The degree of inter-annotator agreement was measured using Cohen's Kappa, which has a value of 0.87, thus denoting strong reliability. This process annotated and distributed 11,320 tweets into 5,109 sarcastic and 6,211 non-sarcastic tweets. A sample of annotated data/ comments from the experimental data is presented in Table 1.

4.4 Data Splitting

The annotated dataset was divided into training, validation, and test sets using a 70:15:15 split ratio. A stratified sampling strategy was employed to ensure that the class distribution of sarcastic and non-sarcastic instances was preserved across all subsets. The validation set was used for hyperparameter tuning and model selection. The final performance results reported in this study are based on evaluation on a held-out test set using a single stratified split. While this provides a consistent evaluation setting, future work will incorporate k-fold cross-validation or repeated holdout evaluation to further assess model robustness.

4.5 Feature Selection

We used combined lexical features to represent the data. These features included a combination of punctuation-based features, n-grams, and Term Frequency–Inverse Document Frequency (TF-IDF) weighting to capture the relative importance of terms within the dataset.

4.6 Text Classification

For the sarcasm classification task, we adopted a set of machine learning algorithms commonly used for text classification due to their complementary learning strategies [10, 18]. This variety of classifiers allows capitalizing on both the strengths of individual models and the robustness of ensemble methods to capture fine-grained linguistic structures of sarcasm in Roman Urdu.

- Support Vector Machines (SVM): A margin-based classifier effective in high-dimensional and sparse feature spaces such as TF-IDF representations.
- Logistic Regression (LR): A probabilistic linear model that estimates class membership using a sigmoid function, commonly used as a strong baseline for short-text classification.
- Decision Trees (DT): A tree-structured model that divides the dataset based on feature thresholds, providing interpretability and capturing non-linear relationships.
- Random Forests (RF): An ensemble of decision trees that improves generalization and reduces variance through aggregated predictions.
- K-Nearest Neighbors (KNN): A distance-based classifier that assigns class labels based on similarity to neighboring instances in the feature space.

In addition to these classifiers, ensemble learning techniques were employed:

Table 1. Annotated Data/Comments from the Experimental Dataset

S.No.	Tweet/Comment	Annotation	Translation
1	Congratulations, uni walon ne sgl band kar di humare batch se	Sarcastic (1)	Congratulations, uni has stopped self-granted leave from our batch
2	Background aur app, beauty he beauty.	Non-Sarcastic (0)	Background and you, beauty and beauty around.
3	Han han bro rat 4 baje uthana parhen gay!!	Sarcastic (1)	Yes yes, bro, get me awake at 4 a.m. in the night, we will study!!
4	Huhh Behn aaj monthiversary hogaye tumhain sochte huay k ye book parthe hoon…..	Sarcastic (1)	Huhh, sis, you have completed a month thinking you will be reading this book…..
5	Yar Koi exam postpone hua he nahen	Non-Sarcastic (0)	No exam has been postponed Yar
6	Ye to dress mujhay bhi chaiye please please	Non-Sarcastic (0)	I also want this dress, please, please
7	Nahe nahe dard kaisa, main insan thore hoon??	Sarcastic (1)	No no, what pain, I am not a human being??
8	Wah yaar, time pe reply denay ka Shukria	Sarcastic (1)	Great yar, thank you for replying on time

- Extreme Gradient Boosting (XGBoost): A boosting algorithm that sequentially combines models to minimize prediction error and has demonstrated strong performance on structured text features.
- Bagging (Bootstrap Aggregating): An ensemble method that trains multiple learners on bootstrapped samples and aggregates their predictions to improve stability and reduce overfitting.

4.7 Performance Evaluation Metrics

To measure the effectiveness of the models, we used the standard metrics including precision, recall, F1-score, and accuracy [24]. Precision measures the reliability of positive predictions, Recall evaluates the ability to identify all true positives, and the F1-score provides a balance between Precision and Recall, particularly useful in imbalanced datasets. Accuracy reflects the overall percentage of correctly classified instances. Moreover, ROC curves were plotted to illustrate the trade-off between true positive and false positive rates, while confusion matrices were analyzed to visualize misclassification patterns and gain deeper insight into classification errors. Fig. 2 presents a summary of the discussed workflow phases.

5 Results and Discussion

This section discusses the results of the proposed technique for sarcasm detection in Roman Urdu text. Since sarcasm has a subtle nature, the connotation can significantly differ from the actual translation, the assessment of models should not be reduced to just accuracy in reporting. To achieve a thorough evaluation, we have implemented and evaluated a variety of seven machine learning models namely Support Vector Machine (SVM), Decision Tree Classifier (DTC), Logistic Regression Classifier (LRC), Random Forest Classifier (RFC), K-Nearest Neighbor Classifier (KNNC), XGBoost Classifier (XGBC), and Bagging Classifier (BGC).

The choice of classifiers was based on the scale of learning paradigms between linear and ensemble-based methods. Their inclusion enables a rigorous comparative analysis and provides complementary perspectives on handling the inherent complexity of sarcasm detection in informal, resource-constrained linguistic settings. Each classifier was selected on the basis of its individual methodological excellence. SVM and Logistic Regression can serve as powerful linear baselines and are well-adapted to high-dimensional sparse text features. Decision Trees are both interpretable and show the non-linear feature interactions prevalent in sarcastic utterances. Random Forest and Bagging, as an ensemble technique, increase stability and generalization to reduce the overfitting challenges likely to be experienced by individual learners. K-Nearest Neighbor presents an instance-based paradigm of classification based on

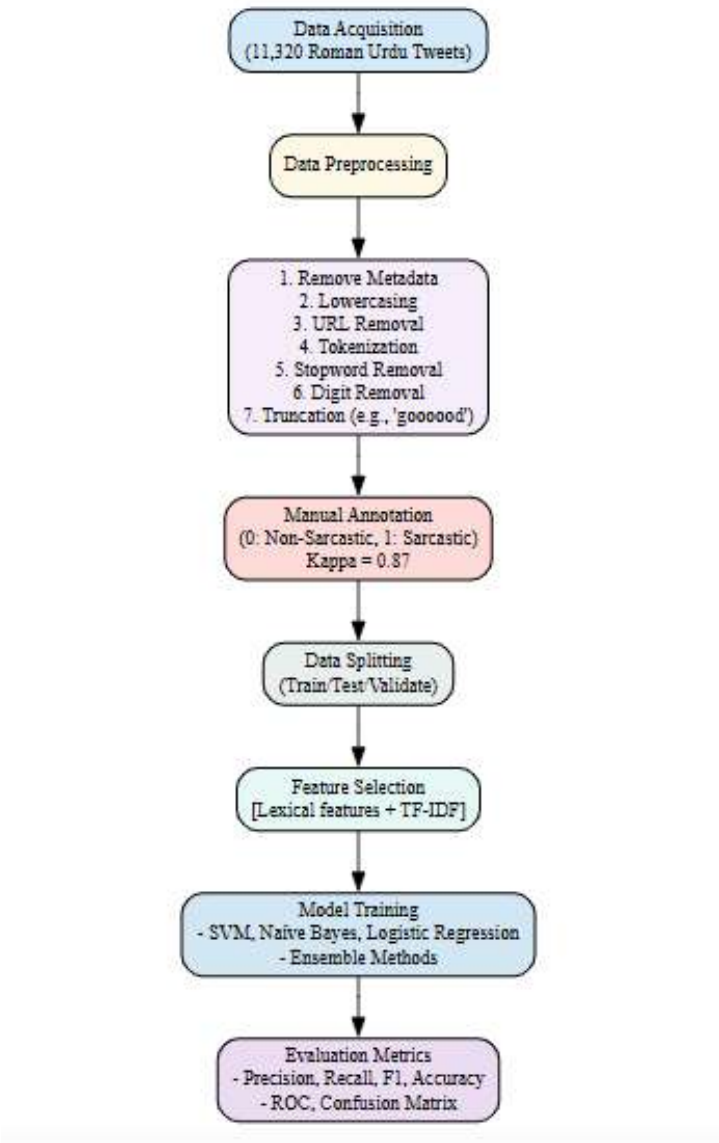


Figure 2. Flowchart for Sarcasm Detection in Roman Urdu

similarity in the feature space, whereas XGBoost is a recent and state-of-the-art boosting algorithm, which is known to provide a means of modeling subtle and complex decision boundaries in a noisy, low-resource environment.

A summary of the comparative performance of these classifiers is presented in Table 2, showing the performance metrics, i.e., accuracy, precision, recall, and F1-scores of each classifier. These findings provide a comprehensive outlook on how the various models can effectively represent the delicate twist of patterns of Roman Urdu text on sarcasm. Random Forest and XGBoost are the best among the models that performed with the highest accuracy of 0.87 and F1-score of 0.88, and thus depict that they are more than capable of learning complex non-linear trends and generalizing them to noisy data. Logistic Regression and SVM are also competitive with an accuracy score of 0.85 and 0.86, respectively, and do not give unbalanced precision-recall trade-offs, making them useful linear baselines. The Bagging Classifier also performed well with an accuracy of 0.86 and the most accurate precision of 0.88 due to the reduction of errors because of the ensemble aggregation. Decision Tree shows marginally lower accuracy of 0.83, but was easy to interpret and gave a level of competitive precision of about 0.87. K-Nearest Neighbor, compared to the other two, shows the least accuracy of 0.74 and the lowest F1-score of 0.79, which reflects poorly on the instance-based learning being less successful in capturing the contextual dependencies of sarcastic expressions.

Table 2. Comparison of the performance results and metrics for implemented techniques

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	0.86	0.85	0.89	0.87
DTC	0.83	0.87	0.81	0.84
LRC	0.85	0.83	0.91	0.87
RFC	0.87	0.85	0.91	0.88
KNNC	0.74	0.72	0.87	0.79
XGBC	0.87	0.85	0.92	0.88
BGC	0.86	0.88	0.86	0.87

A deeper understanding of classifier behavior is revealed through confusion matrix analysis, presented in Figure 3. SVM in Figure 3(a) and Logistic Regression in Figure 3(c) captured sarcastic cues effectively with high true positives but exhibited a tendency to over-predict sarcasm, leading to elevated false positives. The Decision Tree in Figure 3(b) demonstrated competitive precision but suffered from higher false negatives, reducing recall. Random Forest in Figure 3(d) and XGBoost in Figure 3(f) delivered the most balanced outcomes, minimizing both types of misclassifications. Bagging Classifier in Figure 3(g) also achieved strong performance, albeit with slightly more false positives compared to Random Forest and XGBoost. In contrast, K-Nearest Neigh-

bor in Figure 3(e) was the weakest, frequently misclassifying sarcastic utterances as non-sarcastic. Collectively, these results reinforce the robustness of ensemble methods, the reliability of linear baselines, and the limitations of instance-based learning in this context.

The error analysis of implemented models (as shown in Fig. 3) indicates that the models are oriented towards identifying and capturing lexically tied sarcasm instances, which typically use punctuation marks, specific vocabulary patterns or intensifiers. However, semantic irony in which the semantic meaning is usually the opposite of what is said by the user and complex implicit nature of sarcasm still remains an open challenge. Additionally, highly code-mixed patterns were also part of false negatives.

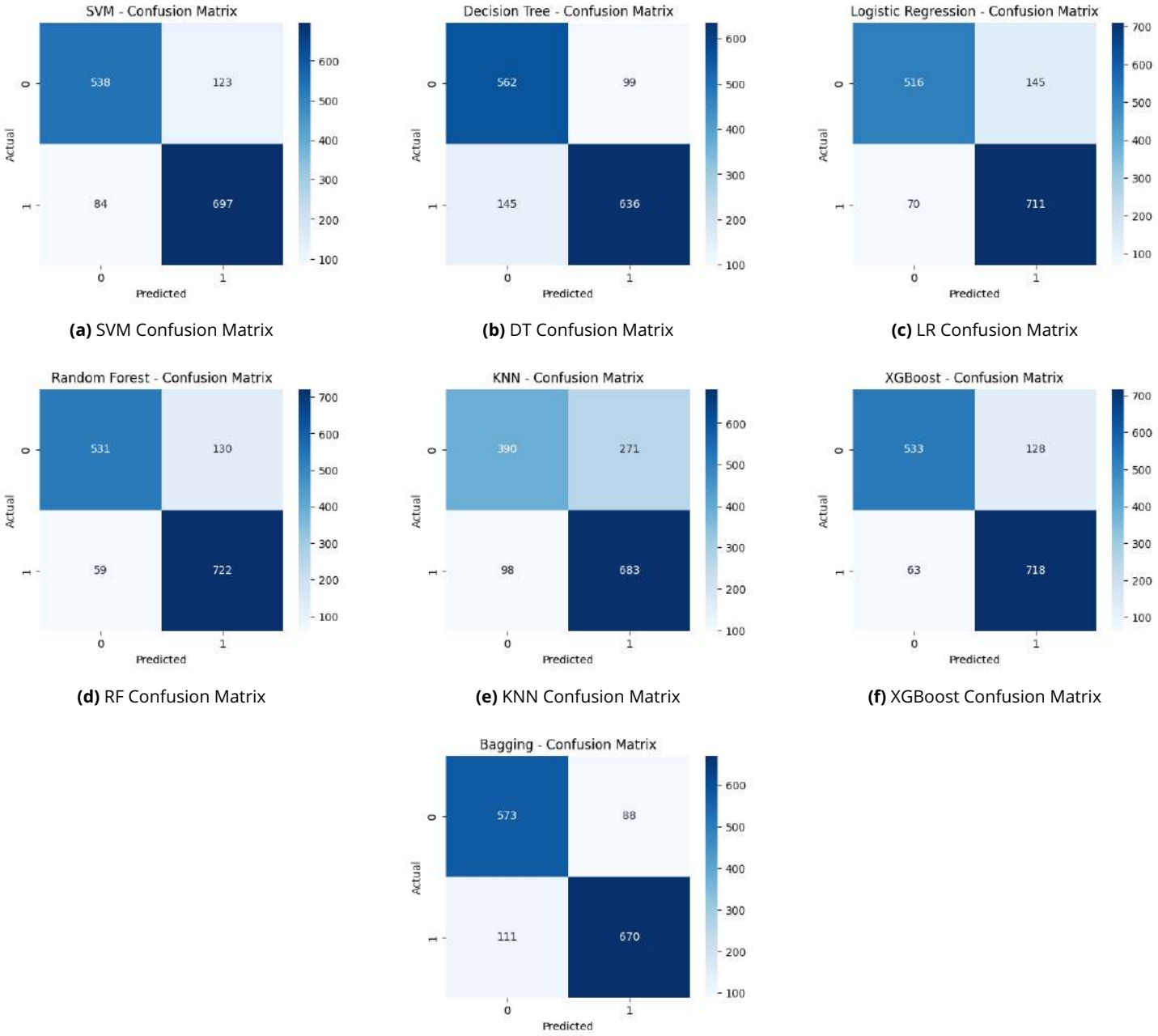
To better understand these models' behavior, a graphical comparison of accuracy, precision, recall, and F1-score amongst classifiers is provided in Figure 4. This has indicated that assigning ensemble-based models, in this case, Random Forest and XGBoost, will result in better outcomes on all metrics, and this confirms their effectiveness in treating complex forms of sarcastic sentences. The bagging classifier also performed rigorously, particularly in aspects of precision. In the meantime, the linear models, including SVM and Logistic Regression, gave competitive baselines, performing well in terms of recall but a little worse in the accuracy metric due to potentially over-predicting tendencies. The Decision Tree continued to perform averagely, its lower accuracy reflected by poorer recall and overall reliability, and K-Nearest Neighbor was the underperformer, as it reflected the lowest performance in all dimensions.

The ROC curve [25] is a statistical approach that evaluates the performance of several types of classification algorithms based on the differentiation between the positive and negative classes, measured by the area under the curve (AUC). The SVM demonstrates an impressive outcome, with an AUC of 0.92, reflecting the high level of classification efficacy, similar to the values of Logistic Regression, Random Forest, XGBoost, and Bagging that reached 0.92, as shown in Figure 5. Comparatively, the Decision Tree model performs averagely with an AUC of 0.83 indicating that it has a less ideal classification capability, and the KNN model performs poorly relative to the other models with an AUC of 0.81, which shows that it has a comparatively weaker classification ability. In general, the SVM, Logistic Regression, Random Forest, XGBoost, and Bagging models with the highest AUC values would be viewed as mainly adequate in predictive modeling of the classification problem in question (since they are closer to the top-left corner of AUC), whereas Decision Tree and KNN would not be considered as well-suited in this area (as their AUC scores are relatively low).

One can say that the empirical results show that ensemble-based techniques, such as Random Forest and XGBoost, work well when it comes to sarcasm detection in the low-resource Roman Urdu data not only in terms of accuracy but also in terms of reasonable precision-recall trade-offs. Other linear models including SVM, and Logistic Regression, are slightly less effective but good baselines since they can be performed in a high-dimensional feature space efficiently. Where Decision Tree may be said to be heralded by interpretability, there is diminished recall performance; K-Nearest Neighbor, on the other hand, provides dissatisfactory performance on evaluation measures. All in all, the findings indicate that ensemble methods are uniquely suited to the problem of detected complexity as well as the noisy aspect of sarcasm in low-resource environments, presenting a bright possibility in future development of novel approaches in computational sarcasm detection.

To position our results with respect to existing work, we compare the performance of the best-performing ensemble model in this study with previously reported results on Urdu/Roman Urdu sarcasm detection. As shown in Table 3, our framework achieves an accuracy of 87% and an F1-score of 0.88 on the constructed Roman Urdu dataset. In contrast, [10] reports a maximum accuracy of 85% using SVM for sarcasm detection in Urdu text, while [4] achieves 79.51% accuracy and an 80.04% F1-score with a multilingual mBERT–BiLSTM–multi-head attention model on an Urdu sarcasm corpus.

These results indicate that the proposed framework establishes a promising baseline for sarcasm detection in Roman Urdu. However, due to differences in datasets, annotation schemes, class distributions, domains, and evaluation settings between prior Urdu studies and our research on Roman-Urdu, direct comparison should be



(a) SVM Confusion Matrix **(b)** DT Confusion Matrix **(c)** LR Confusion Matrix

(d) RF Confusion Matrix **(e)** KNN Confusion Matrix **(f)** XGBoost Confusion Matrix

(g) Bagging Confusion Matrix

Figure 3. Confusion matrices of all implemented classifiers

interpreted with caution.

6 Model Interpretability using XAI and LIME

Although a major objective is to attain high accuracy, at the same time, it is essential to have insights into the reasons a model produces a given prediction to establish trust, guarantee fairness, and optimize the system, particularly in subtle social tasks like the detection of sarcasm. In this regard, we employed Explainable AI (XAI) techniques to interpret the decisions of our best-performing model, which is a Random Forest classifier with an accuracy of 87%.

One of the most popular XAI frameworks is Local

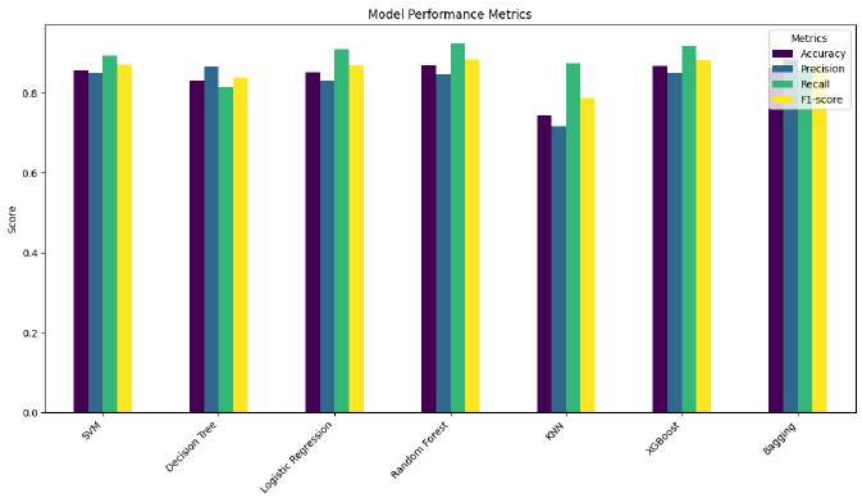


Figure 4. Comparison of classification performance across models using accuracy, precision, recall, and F1-score

Table 3. Comparison with existing Urdu/Roman Urdu sarcasm detection results

Study	Language	Model	Accuracy	F1-Score
[4]	Urdu	mBERT-BiLSTM-MHA	79.51%	80.04%
[10]	Urdu	SVM	85%	84%
This Study	Roman Urdu	Random Forest, XGBoost	87%	88%

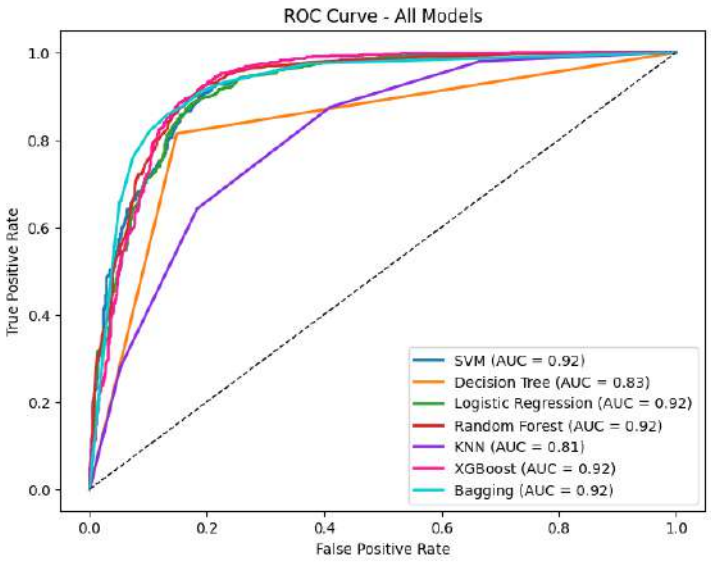


Figure 5. ROC Curve of all the models

Interpretable Model-agnostic Explanations (LIME), which provides explanations of the predictions of any classifier with local approximations based on an interpretable model [26]. To explain the prediction, LIME operates by constructing perturbed copies of the original text instance (e.g., by randomly removing words) and observing the changes in the model's prediction. Relying on such perturbations, LIME produces a linear model faithful to the behavior of the classifier in the locality of the original instance. The result is a list of words or phrases in the input text with a weight indicating its contribution to a particular class (e.g., 'Sarcastic' or 'Not Sarcastic', as in the current study). This provides a human-interpretable explanation for an individual prediction that highlights the essential and key lexical cues that the model relied upon [10].

We applied LIME to analyze predictions on our Roman Urdu test set. The responses continuously showed that our model had learned to link certain words and contextual patterns with sarcasm, which aligns well with the linguistic intuition of native users. For example, consider the following tweet:

- Text: aunty ka jawab dekhen jaag rahe hain bhai ab ye (English translation: Look at auntie's answer, bro she is awake now, huh.)
- True Label: Sarcastic

- Model Prediction: Sarcastic (Probability: 0.72)

This prediction was explained by LIME, as illustrated in Figure 6. The method identifies the words "jawab" (answer), bhai (brother), rahe (are), ab (now), and jaag (awake) as the strongest features to push the model towards the "Sarcastic" category. The fact that these words appear in a certain combination probably indicates to the model an incongruence or excessive formality, which is a characteristic of sarcastic utterances in Roman Urdu. This demonstrates the model's ability to identify context-specific sarcastic expressions, not limited to sentimental words. Similarly, more LIME explanations for Roman Urdu examples executed on the best model are highlighted in Figure 7 and Figure 8. The model correctly classifies Not-Sarcastic and Sarcastic tweets in accordance with their true labels.

In other example instances, LIME revealed that the model correctly discounted common stop-words in Roman Urdu and focused on verbs, nouns, and adjectives that carried semantic weight. This interpretability is crucial in making sure that the model has been reliable in ensuring that it has not learnt a bias or false association. By using LIME, we move beyond the "black box" nature of complex ensemble models towards more explainable open white box interpretations, allowing researchers and prospective end-users to gain insight into the decision-making process of the sarcasm detector with Roman Urdu.

7 Conclusion, Societal Implications, and Future Work Directions

This paper proposes a paradigm to detect sarcasm in Roman Urdu, a low-resource and casual language frequently used on social media. Through the establishment of a manually labeled dataset of 11,320 tweets and the application of machine learning and ensemble approaches, we have performed a systematic study of the performance of different classifiers to detect sarcasm. The findings indicated that models like Random Forest and XGBoost were very effective as compared to single classifiers, with an accuracy of 0.87 and an F1-score of 0.88. Furthermore, the utilization of the LIME technique provided valuable interpretability by showcasing the specific words and phrases that our model leverages to identify sarcasm, thus making predictions more transparent and trustworthy. These findings confirm that ensemble learning is well-suited for capturing the complex linguistic and contextual cues of sarcasm in code-mixed languages like Roman Urdu, which also has orthographic variations and informal constructs as added hurdles. Furthermore, our research presents a robust foundation for enhancing sarcasm detection research in low-resource languages.

The proposed approach has potential implications for society if embedded in different systems and social media platforms, such as strengthening communication interfaces in the digital world, reforming human-computer interaction, fostering online safety, monitoring and improving mental well-being, and fighting deception and misleading information. Furthermore, research in low-resource languages also overcomes the challenges of cultural and linguistic diversities across different communities.

Although the proposed approach performs well, certain limitations should be acknowledged. The reliance on Twitter/X data introduces potential platform-specific and demographic biases, which may limit generalization to other domains such as formal text or conversational platforms. Addressing these aspects could provide a more comprehensive evaluation framework.

Numerous other opportunities also open up in the context of future research extensions. While the current study establishes a solid baseline and validates the efficacy of ensemble learning with lexical features, we recognize the potential of transformer-based architectures in such settings. Incorporating deep learning and transformer-based approaches, such as BERT or UrduBERT, could enhance contextual understanding beyond lexical features. Future work may also focus on analyzing the performance and accuracy gain of mBERT and XLM-RoBERTa language models, specifically fine-tuned on Roman Urdu's complex structure, in contrast to the results of the techniques proposed in this study.

Generalization can also be improved by expanding the dataset with more balanced sarcastic and non-sarcastic examples, including conversational threads. Though conversational threads are quite challenging to capture for new regional languages, due to mixed code communication patterns adopted by users in a single thread. Additionally, other XAI frameworks, like

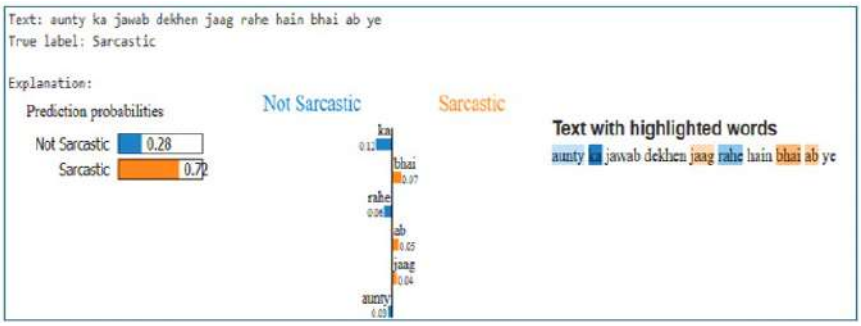


Figure 6. LIME explanation for a correctly classified sarcastic tweet in Roman Urdu (The highlighted words were the most significant contributors to the "Sarcastic" prediction).

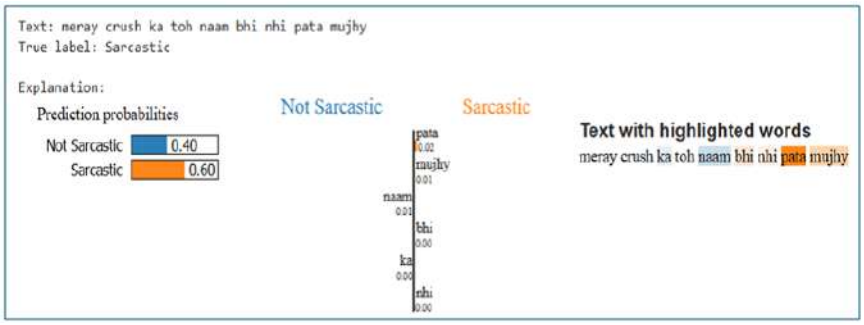


Figure 7. LIME explanation for a correctly classified sarcastic tweet in Roman Urdu (The highlighted words were the most significant contributors to the "Sarcastic" prediction).

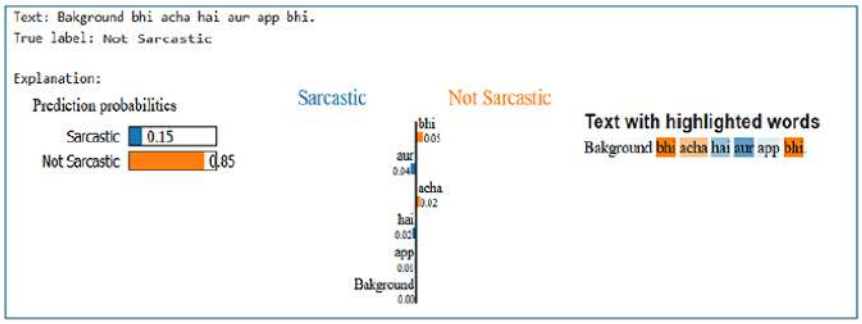


Figure 8. LIME explanation for a correctly classified Not-Sarcastic tweet in Roman Urdu (The highlighted words were the most significant contributors to the "Not Sarcastic" prediction).

SHAP, could also be explored in the future to compare the interpretability of models. Along with that, it can be extended to detect the multimodality of sarcasm by using text and combining it with emojis or visual elements that can potentially provide more insights. Finally, developing real-time sarcasm detection tools for applications in sentiment analysis, opinion mining, and automated moderation would further demonstrate the practical value of this research for society and human well-being.

Acknowledgment

We are extremely grateful to the Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, for providing all the technical support and resources to accomplish this project.

Author Contributions

All the authors of this work contributed equally to study conception, methodology design, implementation, result reporting and validation, draft writing, editing, proofreading, and formatting.

Compliance with Ethical Standards

It is declared that all authors do not have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

AI Assistance Disclosure

The authors declare that the artificial intelligence (AI) tool was used only for minor language editing and refinement after completion of the technical work. No AI tool was used for the generation of research data, analysis, results, interpretations, or cited scholarly content. All AI-assisted content was reviewed and validated by the authors, who take full responsibility for the final manuscript.

Funding Information

This research study received no external funding. All the resources consumed were internally available at the Institute of Information and Communication Technologies, Mehran University of Engineering and Technology, Jamshoro.

References

[1] A. D. Yacoub, S. Slim, and A. Aboutabl, "A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends," *International Journal of Electrical and Computer Engineering Systems*, vol. 15, no. 1, pp. 69–78, 2024.

[2] A. O. Bajeh, A. Shittu, and C. Asiyanbola, "Automatic sarcasm detection in textual data: A literature review," *Sule Lamido University Journal of Science and Technology*, vol. 11, no. 1–2, pp. 233–242, 2025.

[3] A. Mansoori, K. Tahat, O. Al Zoubi, D. N. Tahat, M. Habes, H. Himdi, et al., "Detection of sarcasm in news headlines using NLP and machine learning," in *Generative AI in Creative Industries*, Cham, Switzerland: Springer Nature, 2025, pp. 503–517.

[4] M. E. Hassan, M. Hussain, I. Maab, U. Habib, M. A. Khan, and A. Masood, "Detection of sarcasm in Urdu tweets using deep learning and transformer based hybrid approaches," *IEEE Access*, vol. 12, pp. 61542–61555, 2024.

[5] T. Javed, M. A. Nouman, and R. Zahid, "BERT model adoption for sarcasm detection on Twitter data," *VFAST Transactions on Software Engineering*, vol. 12, no. 3, pp. 177–198, 2024.

[6] N. A. Helal, A. Hassan, N. L. Badr, and Y. M. Afify, "A contextual-based approach for sarcasm detection," *Scientific Reports*, vol. 14, no. 1, p. 15415, 2024.

[7] A. Abuzayed and H. Al-Khalifa, "Sarcasm and sentiment detection in Arabic tweets using BERT-based models and data augmentation," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, 2021, pp. 312–317.

[8] T. Ptáek, I. Habernal, and J. Hong, "Sarcasm detection on Czech and English Twitter," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, Dublin, Ireland, 2014, pp. 213–223.

[9] R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, pp. 13–18, 2023.

[10] S. Khan *et al.*, "An automated approach to identify sarcasm in low-resource language," *PLoS One*, vol. 19, no. 12, p. e0307186, 2024.

[11] M. A. Manzoor, S. Mamoon, S. K. Tao, Z. Ali, M. Adil, and J. Lu, "Lexical variation and sentiment analysis of Roman Urdu sentences with deep neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.

[12] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, and M. Shrivastava, "A corpus of English-Hindi code-mixed tweets for sarcasm detection," *arXiv preprint arXiv:1805.11869*, 2018.

[13] R. Jamil, I. Ashraf, F. Rustam, E. Saad, A. Mehmood, and G. S. Choi, "Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short-term memory network model," *PeerJ Computer Science*, vol. 7, p. e645, 2021.

[14] A. Ashwitha, G. Shruthi, H. R. Shruthi, and T. C. Manjunath, "Sarcasm detection in natural language processing," *Materials Today: Proceedings*, vol. 37, pp. 3324–3331, 2021.

[15] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Context-aware sarcasm detection using BERT," in *Proceedings of the Second Workshop on Figurative Language Processing*, Association for Computational Linguistics, 2020, pp. 83–87.

[16] S. K. Lora, I. Jahan, R. Hussain, R. Shahriyar, and A. B. M. A. Al Islam, "A transformer-based generative adversarial learning to detect sarcasm from Bengali text with correct classification of confusing text," *Heliyon*, vol. 9, no. 12, 2023.

[17] S. Sengupta, S. Ghosh, P. Mitra, and T. I. Tamiti, "Milestones in Bengali sentiment analysis leveraging transformer-models: Fundamentals, challenges and future directions," *arXiv preprint arXiv:2401.07847*, 2024.

[18] S. K. Lora *et al.*, "Ben-Sarc: A self-annotated corpus for sarcasm detection from Bengali social media comments and its baseline evaluation," *Natural Language Processing*, vol. 31, no. 2, pp. 674–699, 2025.

[19] M. A. Hasan *et al.*, "Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis," in *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024, pp. 17808–17818.

[20] S. Chanda, A. Mishra, and S. Pal, "Sarcasm detection in Tamil and Malayalam Dravidian code-mixed text," in *Proceedings of FIRE (Working Notes)*, 2023, pp. 336–343.

[21] A. Kirkpatrick and I. Schaller-Schwaner, "English as a lingua franca," in *Handbook of Practical Second Language Teaching and Learning*, Routledge, 2022, pp. 97–113.

[22] S. Dhall, S. Kumar, and S. Kumar, "A review on sentiment analysis in low-resource languages focusing on fake news and sarcasm detection as major challenges," *SN Computer Science*, vol. 6, no. 6, p. 693, 2025.

[23] A. D. Yacoub, A. E. Aboutabl, and S. Slim, "A survey of challenges, methods, and trends in sentiment analysis and sarcasm detection," *FCI-H Informatics Bulletin*, vol. 6, no. 2, pp. 61–68, 2024.

[24] D. Šandor and M. Bagić Babac, "Sarcasm detection in online comments using machine learning," *Information Discovery and Delivery*, vol. 52, no. 2, pp. 213–226, 2024.

[25] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Proceedings of the Computer Science On-line Conference*, Springer, 2023, pp. 15–25.

[26] T. Aljrees, "Improving prediction of Arabic fake news using ELMO's features-based tri-ensemble model and LIME XAI," *IEEE Access*, vol. 12, pp. 63066–63076, 2024.