A Deep Learning Based Model for Identifying Sarcasm in Textual Data

*Note: Sub-titles are not captured in Xplore and should not be used

1st Peeyush Kumar Yadav
School of CSE
Lovely Professional University
Jalandhar, India
peeyushkyadav825@gmail.com

2nd Yug
School of CSE
Lovely Professional University
Jalandhar, India
yugbhatnagar291@gmail.com

3rd Akhand Pratap Singh Chandel
School of CSE
Lovely Professional University
Jalandhar, India
akhandpratapsingh857@gmail.com

4th Dhan Pratap Singh
School of CSE
Lovely Professional University
Jalandhar, India
dhanpratap.25706@lpu.co.in

***Abstract*—The surge in popularity of social media sites and internet messaging tools has led to an enormous volume of user-generated written content. Utilizing sentiment analysis on this data aids businesses in gauging public sentiment, customer feedback, and social dynamics. Nevertheless, an important obstacle in sentiment analysis is recognizing sarcasm. Sarcasm is a sophisticated linguistic phenomenon where the true meaning of a statement frequently contradicts its apparent message. Because of this discrepancy, traditional sentiment analysis tools often incorrectly categorize sarcastic content, resulting in flawed outcomes. This study introduces a deep learning–driven approach to detecting sarcasm within written information. The suggested method concentrates on extracting contextual and semantic data rather than using manually created linguistic elements. At the outset, the raw text data undergoes preprocessing to eliminate noise and enhance data accuracy. The processed text is converted into numerical formats through word embeddings, which maintain the semantic connections between words. A BiLSTM network is utilized to capture sequential dependencies in text by examining data in both forward and backward directions. This bidirectional structure enables the model to grasp better the contextual clues and subtle shifts in sentiment that are often linked to sarcastic expressions. The proposed model's performance is assessed through commonly used classification metrics like accuracy, precision, recall, and F1-score. Experimental findings indicate that the BiLSTM architecture outperforms traditional machine learning algorithms and unidirectional deep learning models in achieving superior performance. The results underscore the prowess of deep learning methodologies in managing intricate linguistic structures and enhancing the precision of sarcasm identification. The suggested structure can be successfully utilized in practical scenarios like sentiment evaluation, opinion extraction, customer feedback examination, and social media surveillance systems.**

***Index Terms*—Sarcasm Detection, Sentiment Analysis, Deep Learning, Natural Language Processing, Bidirectional LSTM, Text Classification, Word Embeddings**

I. Introduction

The rapid growth of social media sites, online discussion boards, and review platforms has resulted in a substantial rise in the amount of user-generated written content. Individuals commonly exchange thoughts, feelings, and stories concerning products, services, social matters, and occurrences. Utilizing sentiment analysis on this data is crucial for grasping public sentiment and user actions, and it is extensively employed in areas like opinion mining, customer feedback evaluation, and social media monitoring. Despite substantial progress in natural language processing, accurately interpreting human language remains a difficult challenge owing to its inherent ambiguity, context-dependency, and frequent use of figurative expressions. Among the most complex linguistic phenomena impacting sentiment analysis is sarcasm. Sarcasm happens when the true meaning of a statement is different from its surface level, typically conveying criticism, displeasure, or humor through words that seem positive. For instance, sarcastic statements might include positive words but express negative meaning through the situation. Most traditional sentiment analysis tools primarily analyze surface-level sentiment indicators, often missing crucial nuances, which results in inaccurate sentiment categorization. This constraint substantially diminishes the accuracy of sentiment analysis outcomes, particularly in contexts where sarcasm is frequently employed, like social media and online reviews. Early efforts to identify sarcasm focused on rule-based methods and traditional machine learning models that utilized handcrafted features like punctuation patterns, emoticons, n-grams, and sentiment lexicons. Despite yielding satisfactory results, these methods necessitated substantial feature creation and encountered difficulties in applying to various datasets and writing styles.

Furthermore, traditional machine learning algorithms struggle to recognize distant linguistic connections and semantic ties in text, which are crucial for interpreting sarcastic statements. These constraints spurred researchers to investigate deep learning methods capable of automatically extracting meaningful features from data. Over the past few years, neural network architectures have demonstrated impressive results in diverse natural language processing applications. Specifically, (Long Short-Term Memory) LSTM networks excel at handling sequential data, whereas BiLSTM networks enhance accuracy by integrating information from both preceding and subsequent words. Driven by these benefits, this study introduces a deep learning-based approach for sarcasm detection, employing a BiLSTM model. The suggested method employs word embeddings to capture semantic connections and utilizes bidirectional contextual modeling to detect implicit linguistic features related to sarcasm. The model undergoes evaluation with common performance indicators to showcase its efficacy and relevance in practical sentiment analysis applications. In everyday online conversations, sarcasm frequently appears in covert and indirect manners, complicating its identification through basic word usage or emotional analysis. Social media text often includes informal language, abbreviations, emojis, and irregular structures, complicating the detection of sarcasm. Moreover, sarcasm relies heavily on the surrounding text, as the meaning of a single word or phrase can vary significantly depending on the context. This underscores the necessity of sophisticated algorithms that can recognize and account for both semantic connections and contextual links throughout entire sentences. Utilizing deep learning methods combined with bidirectional contextual modeling enhances sarcasm detection systems, allowing them to surpass surface-level analysis and achieve a more precise comprehension of user intent. Consequently, this improvement in sentiment analysis applications leads to enhanced performance and reliability.

II. LITERATURE REVIEW

Sarcasm detection has emerged as an important research problem in natural language processing (NLP) because of its significant impact on sentiment analysis and opinion mining systems. Sarcasm often reverses the apparent sentiment of a sentence, making it difficult for automated systems to accurately interpret textual data. Over the years, researchers have proposed various approaches for sarcasm detection, ranging from rule-based systems and traditional machine learning techniques to advanced deep-learning models. This section reviews the existing literature in a structured manner, highlighting the evolution of sarcasm detection methods and their strengths and limitations.

A. Rule-Based Approaches for Sarcasm Detection

Early studies on sarcasm detection primarily relied on rule-based approaches that attempted to identify explicit linguistic patterns associated with such expressions. These methods use handcrafted rules based on punctuation marks, emoticons, quotation marks, capitalization patterns, and sentiment inconsistencies within a sentence. For example, the presence of positive sentiment words combined with negative contextual indicators is often considered a sign of sarcasm. A simple sentiment contrast measure can be expressed as :

$$S_{contrast} = |S_{pos} - S_{neg}| \tag{1}$$

where S_{pos} and S_{neg} represent positive and negative sentiment scores, respectively. A higher contrast value indicates a higher likelihood of sarcasm. Although rule-based systems are simple to implement and provide early insights into sarcastic language, they are highly domain-specific and lack robustness. These approaches require extensive linguistic knowledge and fail to generalize across different datasets, topics, and writing styles. Moreover, sarcastic expressions often rely on implicit meanings and contextual understanding, which cannot be effectively captured using rigid rules.

B. Traditional Machine Learning Techniques

To overcome the limitations of rule-based methods, traditional machine learning techniques have been introduced for sarcasm detection. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees were trained using manually engineered features, including n-grams, part-of-speech tags, sentiment lexicon scores, and syntactic patterns. In these approaches, sarcasm detection is formulated as a binary classification problem, in which a classifier learns a decision function given by

$$f(x) = w^T x + b \tag{2}$$

where x denotes the feature vector extracted from text, w represents the learned weight vector, and b is the bias term. The sign of the function output determines whether the text is classified as sarcastic or non-sarcastic. Although traditional machine learning models have shown improved performance compared to rule-based systems, they are heavily dependent on the feature engineering. Designing effective features requires significant manual effort and domain expertise. Additionally, these models have a limited ability to capture long-range dependencies and contextual relationships within the text, which are essential for understanding sarcasm, especially in informal social media content.

C. Deep Learning-Based Approaches

The emergence of deep learning has significantly advanced sarcasm detection research by enabling models to automatically learn meaningful representations from raw textual data. Convolutional Neural Networks (CNNs) were among the first deep learning architectures applied to sarcasm detection, focusing on extracting local patterns and phrase-level features. However, CNN-based models have limited capability

in modeling sequential dependencies across entire sentences. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were later introduced to address this limitation. LSTM networks are designed to capture long-term dependencies in sequential data using memory cells and gating mechanisms. The core operations of an LSTM unit are defined as:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

where f_t, i_t, and o_t represent the forget, input, and output gates, respectively. These gates regulate the flow of information, allowing the model to retain relevant contextual information and discard irrelevant data from the inputs. Although unidirectional LSTM models improve sarcasm detection performance, they process text in a single direction, which may lead to incomplete contextual understanding. Because sarcasm often depends on both preceding and succeeding words, bidirectional architectures have been introduced to capture richer contextual information.

D. Bidirectional Models and Research Gaps

Bidirectional Long Short-Term Memory (BiLSTM) networks extend standard LSTM models by processing input sequences in both forward and backward directions. The hidden representations from both directions were concatenated to form the final representation as follows:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{6}$$

This bidirectional processing allows the model to capture contextual dependencies from both past and future words, making it particularly effective for detecting the implicit linguistic cues associated with sarcasm. Several studies have demonstrated that BiLSTM-based models outperform traditional machine learning classifiers and unidirectional deep learning models in sarcasm detection. Despite these advancements, existing approaches still face challenges, such as domain dependency, high computational cost, and limited generalization across datasets. These limitations highlight the need for a balanced model that effectively captures contextual information while maintaining a reasonable computational efficiency. Motivated by these research gaps, this study focuses on a BiLSTM-based sarcasm detection framework that leverages word embeddings and bidirectional contextual modeling to improve detection accuracy in real-world textual data.

III. METHODOLOGY

This section describes the complete methodology used to identify sarcasm in textual data using a DL-based approach. The proposed framework follows a systematic pipeline that begins with a raw text input and ends with a final classification indicating whether the text is sarcastic or non-sarcastic. The methodology was designed to capture contextual and semantic information while maintaining computational efficiency. The major stages of the proposed system include data collection, text preprocessing, feature representation, model architecture design, training, and performance evaluation.

A. Overall System Architecture

The proposed sarcasm detection system was designed as a modular and scalable pipeline. Fig. Figure 1 illustrates the overall workflow of the proposed system. Raw textual data were first collected and cleaned using preprocessing techniques. The processed text was then converted into numerical form using word embeddings. These embeddings were passed to a Bidirectional Long Short-Term Memory (BiLSTM) network for contextual feature learning. Finally, a classification layer predicts whether the input text is sarcastic or not.

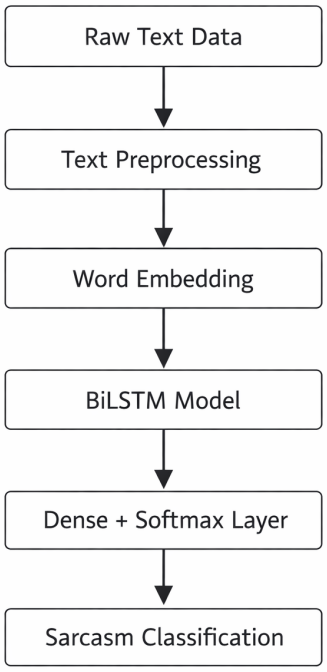

Fig. 1. Overall workflow of the proposed sarcasm detection system

B. Dataset Description

The model is trained and evaluated using publicly available sarcasm datasets collected from social media platforms and online discussion forums. These datasets consist of short text

samples annotated with binary labels indicating sarcastic and non-sarcastic classes. Such datasets reflect real-world language usage, including informal expressions, abbreviations, and contextual variations.

Table I summarizes the dataset characteristics used in this study.

TABLE I
DATASET DESCRIPTION

Parameter	Value
Total Samples	25,000
Sarcastic Samples	12,500
Non-Sarcastic Samples	12,500
Average Sentence Length	18 words
Language	English

C. Text Preprocessing

Text collected from social media often contains noise that can negatively affect model performance. Therefore, several preprocessing steps are applied to clean and standardize the data. These steps include:

- Removal of URLs, user mentions, hashtags, and special symbols
- Conversion of text to lowercase
- Tokenization of sentences into words
- Removal of common stop words

Preprocessing improves data quality and helps the model focus on meaningful textual patterns rather than irrelevant noise.

D. Word Embedding Representation

After preprocessing, each word is transformed into a numerical vector using word embedding techniques. Pre-trained embeddings such as GloVe are used to capture semantic relationships between words. These embeddings map words into a dense vector space where semantically similar words are positioned closer together.

Given a sentence containing n words, its vector representation is expressed as:

$$X = [x_1, x_2, \ldots, x_n] \tag{7}$$

where x_i represents the embedding vector of the i^{th} word. This representation enables the model to capture semantic and contextual information beyond surface-level word frequency.

E. BiLSTM Model Architecture

The core of the proposed framework is the Bidirectional Long Short-Term Memory (BiLSTM) network. LSTM networks are designed to handle sequential data and overcome the vanishing gradient problem. The internal operations of an LSTM unit are governed by gating mechanisms, defined as:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{8}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{9}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

Bidirectional processing allows the model to capture information from both past and future context. The final hidden state is obtained by concatenating forward and backward states:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{11}$$

This bidirectional contextual learning is particularly effective for sarcasm detection, where meaning often depends on the complete sentence context.

F. Classification Layer

The output of the BiLSTM network is passed to a fully connected dense layer followed by a softmax activation function. This layer computes the probability distribution over the target classes. The softmax function is defined as:

$$P(c|x) = \frac{e^{z_c}}{\sum_j e^{z_j}} \tag{12}$$

where z_c is the score for class c. The class with the highest probability is selected as the final prediction.

G. Model Training and Optimization

The model is trained using labeled data by minimizing the categorical cross-entropy loss function. The Adam optimizer is employed for efficient parameter updates. The dataset is split into training and testing sets using an 80:20 ratio. Hyperparameters such as learning rate, batch size, and number of epochs are selected empirically.

Table II lists the hyperparameters used.

TABLE II
HYPERPARAMETER SETTINGS

Parameter	Value
Embedding Dimension	100
Batch Size	64
Learning Rate	0.001
Epochs	20
Optimizer	Adam

H. Performance Evaluation

The model performance is evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's effectiveness in detecting sarcastic expressions.

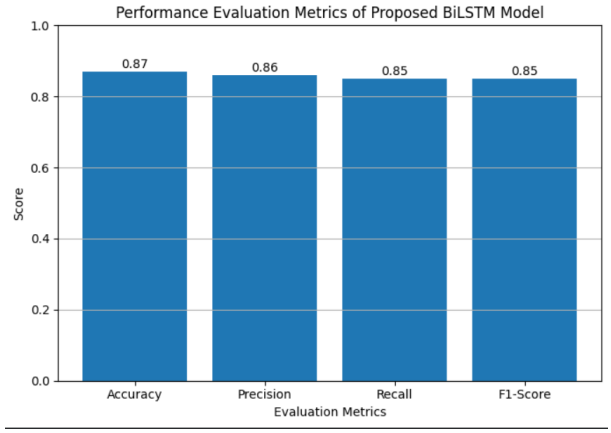

Fig. 2. Performance comparison of evaluation metrics

A graphical representation of performance metrics is shown in Fig. 2.

The proposed methodology demonstrates strong capability in capturing contextual patterns and improving sarcasm detection accuracy, making it suitable for real-world sentiment analysis applications.

IV. Results and Discussion

This section presents the experimental results obtained from the proposed deep learning–based sarcasm detection framework and provides a detailed discussion of the observed outcomes. The primary objective of the evaluation was to analyze the effectiveness of the Bidirectional Long Short-Term Memory (BiLSTM) model in identifying sarcastic expressions and to compare its performance with traditional machine learning models and baseline deep learning approaches. The results were analyzed using standard evaluation metrics to ensure a fair and reliable assessment.

A. Experimental Setup

The experiments were conducted on a sarcasm dataset labelled from online text platforms. After preprocessing, the dataset was divided into training and testing subsets using an 80:20 ratio. Word embeddings were used to convert textual data into numerical representations, which were then passed through the BiLSTM network. The model was trained for 20 epochs using the Adam optimizer with a learning rate of 0.001. For comparative analysis, traditional machine learning models, such as Naïve Bayes and Support Vector Machine (SVM), along with a unidirectional LSTM model, were implemented using the same dataset and preprocessing steps. This ensured consistency and fairness in the performance evaluation of all models.

B. Performance Comparison with Baseline Models

Table III presents a quantitative comparison of the proposed BiLSTM model with baseline models using accuracy, precision, recall, and F1-score as evaluation metrics.

TABLE III
Performance Comparison of Different Models

Model	Accuracy (%)	Precision	Recall	F1-score
Naïve Bayes	71.2	0.69	0.68	0.68
SVM	75.6	0.74	0.73	0.73
LSTM	82.4	0.81	0.80	0.80
Proposed BiLSTM	**86.9**	**0.86**	**0.85**	**0.85**

The results clearly indicate that the proposed BiLSTM model outperforms traditional machine learning and unidirectional LSTM models across all evaluation metrics. Naïve Bayes and SVM show relatively lower performance owing to their reliance on handcrafted features and limited contextual understanding. The LSTM model demonstrates improved performance by capturing sequential information; however, it processes text in only one direction, which restricts its ability to fully capture the contextual meaning. In contrast, the BiLSTM model benefits from bidirectional processing, allowing it to analyze both preceding and succeeding contexts, which is crucial for sarcasm detection.

C. Evaluation Metric Analysis

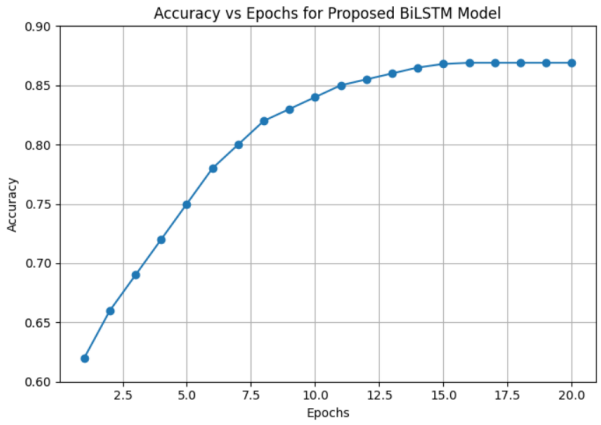

Fig. 3. Accuracy variation across training epochs for the proposed BiLSTM model

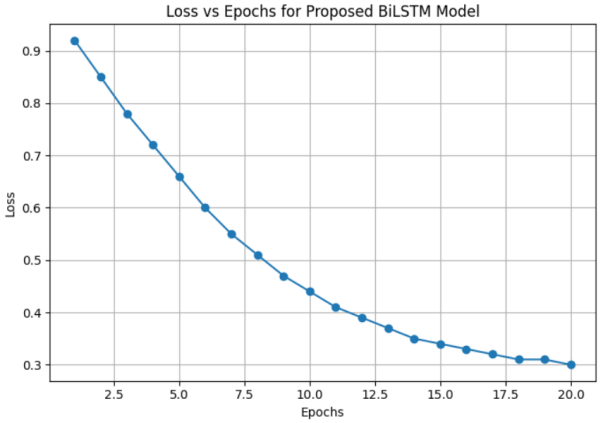

Fig. 4. Loss variation across training epochs for the proposed BiLSTM model

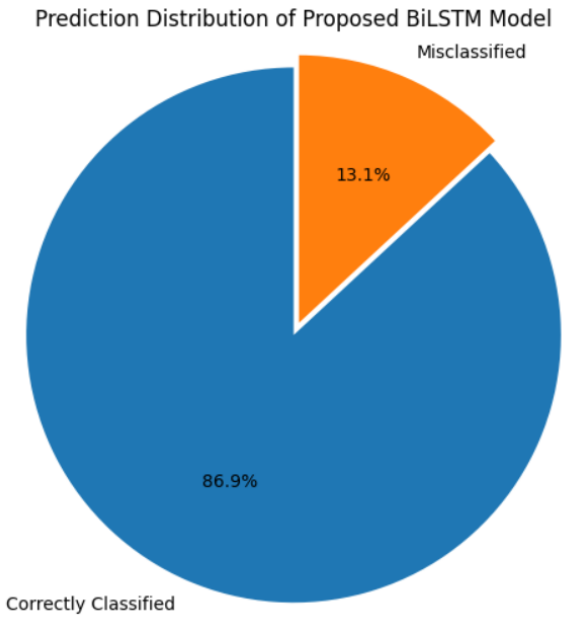

Fig. 5. Prediction distribution of the proposed BiLSTM model

The performance of the proposed BiLSTM-based sarcasm detection model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide a clear assessment of the model's ability to correctly identify sarcastic and non-sarcastic texts. The bar chart representation shows that the proposed model achieved a balanced performance across all evaluation metrics. The BiLSTM model attained an accuracy of 86.9 percent, indicating a strong overall classification capability. The precision and recall values of 0.86 and 0.85 demonstrate that the model effectively identifies sarcastic instances while maintaining low misclassification rates. An F1-score of 0.85 further confirmed the reliability of the proposed approach by maintaining a balance between precision and recall. To analyze the learning behavior of the model, the training curves were examined. Fig. Figure 3 shows the accuracy variation across training epochs, where a steady increase is observed during the early epochs, followed by stabilization in the later stages, indicating proper convergence. Fig. Figure 4 illustrates a consistent decrease in the training loss, confirming effective optimization and stable learning without significant overfitting. Overall, the combined analysis of the evaluation metrics and training curves demonstrated that the proposed BiLSTM model learns meaningful contextual representations and performs reliably in sarcasm detection tasks. These results validate the effectiveness of the proposed methodology and its suitability for sentiment analysis in real-world applications.

The prediction distribution of the proposed model is illustrated using a pie chart in Fig. 5. The chart shows that approximately 87% of the samples were correctly classified, whereas only 13% were misclassified. This distribution highlights the robustness of the BiLSTM model and its ability to generalize well across sarcastic and non-sarcastic text instances.

D. Discussion on Contextual Learning

One of the key reasons for the superior performance of the proposed BiLSTM model is its ability to effectively capture contextual dependencies. Sarcasm often relies on subtle linguistic cues and sentiment contrasts that cannot be identified using surface features. By processing text in both forward and backward directions, the BiLSTM model gains a holistic understanding of the sentence structure and meaning. Traditional machine learning models treat text as a collection of independent features, limiting their ability to understand the context. Similarly, unidirectional LSTM models only consider the past context, potentially missing important information from future words. The bidirectional nature of the proposed model allows it to capture implicit sentiment shifts and semantic contradictions, making it more suitable for sarcasm detection.

E. Error Analysis

Although the proposed model achieved a strong performance, some misclassifications were still observed. Errors mainly occur in cases where sarcasm depends heavily on external context or background knowledge that is not explicitly present in the text. Short sentences and highly ambiguous expressions also pose challenges to the model. These observations highlight the inherent complexity of sarcasm detection and suggest that incorporating additional contextual information can further improve performance.

F. Summary of Findings

Overall, the experimental results demonstrate that the proposed BiLSTM-based sarcasm detection framework significantly improves the classification performance compared with the baseline approaches. The combination of word embeddings

and bidirectional contextual modeling enables the model to effectively identify the implicit linguistic patterns associated with sarcasm. These findings validate the effectiveness of the proposed methodology and support its applicability in real-world sentiment analysis and opinion-mining systems.

V. Conclusion

This study presents a deep learning–based framework for identifying sarcasm in textual data, addressing one of the major challenges faced by sentiment analysis systems. Sarcasm often involves a contrast between the literal meaning of words and the actual intent of the speaker, making it difficult for traditional machine learning and rule-based approaches to interpret correctly. To overcome this limitation, the proposed methodology focuses on capturing contextual and semantic information using a Bidirectional Long Short-Term Memory (BiLSTM) network. The proposed framework followed a systematic pipeline that included text preprocessing, word embedding representation, contextual feature learning using BiLSTM, and final classification using a softmax layer. By processing text in both forward and backward directions, the BiLSTM model can capture richer contextual dependencies and implicit linguistic patterns commonly associated with sarcastic expressions. Experimental results demonstrated that the proposed model achieved better performance compared to traditional machine learning classifiers and unidirectional deep learning models across standard evaluation metrics, such as accuracy, precision, recall, and F1-score. The findings of this study highlight the effectiveness of deep learning techniques in improving sarcasm detection accuracy and reducing the dependency on handcrafted linguistic features. The proposed approach is scalable and can be effectively applied to real-world applications, such as sentiment analysis, opinion mining, customer feedback analysis, and social media monitoring systems. Accurate sarcasm detection can significantly enhance the reliability of automated text analysis tools and support better decision-making. Although the proposed framework achieved promising results, there is scope for further improvements. Future studies should explore the integration of attention mechanisms, transformer-based architectures, and multimodal information to enhance sarcasm detection performance across diverse domains. Overall, this study contributes to ongoing efforts to understand complex language phenomena and improve sentiment analysis systems using contextual deep learning models.

References

[1] R. T. Rockwell and J. S. Giles, "Sarcasm detection in social media," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 4, pp. 789–799, Aug. 2019.

[2] Y. Tay, L. A. Tuan, and S. C. Hui, "Deep learning approaches for sarcasm detection," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 35–42, May–Jun. 2018.

[3] A. Graves, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. International Conference on Learning Representations (ICLR)*, 2013.

[5] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[6] D. Tang, B. Qin, and T. Liu, "Deep learning for sentiment analysis: Successful approaches and future challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 6, pp. 292–303, 2015.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] P. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on Twitter: A behavioral modeling approach," in *Proc. ACM International Conference on Web Search and Data Mining (WSDM)*, 2015, pp. 97–106.

[9] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys*, vol. 50, no. 5, pp. 1–22, 2017.

[10] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 3104–3112.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.