iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VII Month of publication: July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84073

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Enhanced Sentiment Analysis Using RoBERTa and BiLSTM: A Context-Aware Hybrid Deep Learning Approach

Dr. Veguru Gayatri[1], Dr. Rajani Rajalingam[2]
[1]Associate Professor, [2]Professor, Department of Computer Science and Engineering, Geethanjali Institute of Science and Technology (Autonomous), Nellore, Andhra Pradesh, India

***Abstract:** With the widespread growth of digital platforms, online interaction has become an essential part of everyday life. Users frequently express their opinions, feedback, and emotions through reviews and comments on various platforms. Analyzing such textual data plays a crucial role in understanding user sentiment and supporting effective decision-making. However, sentiment analysis faces several challenges, including long-range dependencies within text and the presence of unknown words and symbols. Traditional sentiment analysis approaches mainly rely on sequential models, which process text step by step and often require higher computational time. In contrast, Transformer-based models offer improved efficiency through parallel processing. To address these challenges, this paper presents a context-aware hybrid deep learning approach by integrating the Robustly Optimized BERT Pretraining Approach (RoBERTa) with Bidirectional Long Short-Term Memory (BiLSTM) networks. RoBERTa is employed to generate rich contextual word embeddings, while BiLSTM captures long-term semantic dependencies by processing text in both forward and backward directions. The proposed model is trained and evaluated on the Twitter US Airline Sentiment dataset comprising 14,299 samples across three sentiment classes. Experimental analysis demonstrates that the hybrid approach achieves an accuracy of 85.14% and an F1-score of 0.8487, highlighting its effectiveness for sentiment analysis tasks compared to baseline models.*
***Keywords:** Sentiment Analysis, RoBERTa, BiLSTM, Natural Language Processing, Deep Learning, Transformer, Hybrid Model, Twitter Dataset.*

I. INTRODUCTION

In recent years, the rapid growth of digital platforms and social media has led to a massive increase in user-generated content such as reviews, comments, and feedback. These textual data sources contain valuable information about public opinions and user sentiments. Sentiment Analysis, a subfield of Natural Language Processing (NLP), is used to analyze such text and classify it into categories such as positive, negative, and neutral [1]. Sentiment analysis plays a crucial role in various domains including business, marketing, customer service, and social media monitoring. Organizations leverage sentiment analysis to understand customer opinions, improve services, and make data-driven decisions [4].

Traditional sentiment analysis techniques often fail to capture the true meaning of text due to complex sentence structures, sarcasm, and contextual variations [8]. Sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks process text step by step, which increases computational time and limits their ability to handle long-range dependencies effectively [5]. In contrast, Transformer-based models such as BERT and RoBERTa offer improved efficiency through parallel processing and self-attention mechanisms [9]. However, standalone Transformer models may not fully capture sequential dependencies present in textual data.

To address these limitations, this paper presents an enhanced sentiment analysis system using a hybrid deep learning approach that combines RoBERTa (Robustly Optimized BERT Pretraining Approach) and BiLSTM (Bidirectional Long Short-Term Memory). RoBERTa is employed to generate contextual word embeddings that capture the semantic meaning of words based on their surrounding context, while BiLSTM captures the sequential dependencies in the text by processing it in both forward and backward directions [1]. The proposed hybrid model is trained and evaluated using the Twitter US Airline Sentiment dataset and is designed to effectively classify text into positive, negative, and neutral categories.

The key challenges addressed in this work include the difficulty in capturing contextual meaning of words, the inability of traditional models to handle long-range dependencies, the presence of noisy data such as symbols, slang, and informal language, and the reduced accuracy of existing models in handling real-world social media data [4], [8]. The primary objective of this paper is to develop an efficient and accurate sentiment classification system by integrating the strengths of Transformer-based contextual understanding and recurrent sequence learning.

The remainder of this paper is organized as follows. Section II presents a review of related work. Section III describes the proposed methodology in detail. Section IV discusses the experimental setup and results. Section V concludes the paper and outlines future directions.

II. RELATED WORK

Sentiment analysis has been widely studied using various machine learning and deep learning techniques. This section reviews key contributions in the field that are relevant to the proposed approach.

Rahman et al. [1] proposed a hybrid deep learning model combining RoBERTa and BiLSTM for sentiment analysis. Their approach utilized RoBERTa for generating contextual word embeddings and BiLSTM for capturing sequential dependencies. The model was evaluated on datasets such as IMDb, Twitter US Airline, and Sentiment140, and the results demonstrated that the proposed model outperformed existing methods, achieving higher accuracy and F1-scores.

Sokolová et al. [2] investigated sentiment analysis using Transformer models including BERT, GPT, and T5. Their study integrated BERT-based contextual embeddings with a BiLSTM network to handle complex and noisy social media text. The model incorporated an attention mechanism to focus on important sentiment-bearing words and achieved F1-scores above 0.94 across multiple sentiment classes.

Zhang et al. [3] presented a BERT-LSTM-Attention framework for robust multi-class sentiment analysis on Twitter data. Their model combined BERT for contextual embedding generation with a BiLSTM network and a custom attention mechanism. The experimental results demonstrated that the proposed model outperformed baseline models such as standalone BERT and BERT combined with BiLSTM, achieving high classification performance across all sentiment categories.

Bhonde and Prasad [4] provided a comprehensive overview of sentiment analysis methods, applications, and challenges. Their study discussed various feature selection techniques and classification approaches based on machine learning and NLP. The work highlighted important applications such as business intelligence, product recommendation systems, and political analysis, while addressing challenges such as handling large-scale data and improving classification accuracy.

Younesi et al. [5] proposed a CNN-BiLSTM-based deep learning model for sentiment classification. The model was evaluated using the SemEval-2017 Task 4 dataset and addressed issues such as data imbalance through data augmentation techniques. The experimental results showed that the proposed model achieved an accuracy of approximately 84%, demonstrating the effectiveness of hybrid architectures in enhancing sentiment analysis performance.

Pandey and Singh [6] developed a BERT-LSTM model for sarcasm detection in code-mixed social media posts, highlighting the challenges of processing informal and multilingual content. Tam et al. [7] proposed a ConvBiLSTM deep learning model for Twitter sentiment classification, demonstrating the potential of combining convolutional and recurrent architectures. Alrumaih et al. [8] studied sentiment analysis of comments in social media using various approaches. Vaswani et al. [9] introduced the Transformer architecture, which laid the foundation for subsequent models such as BERT and RoBERTa.

Several other studies have explored sentiment analysis using BERT-based models [10], [11], machine learning classifiers [12], [13], and deep learning architectures [14], [15]. Recent surveys have examined aspect-based sentiment analysis [17], deep learning-based sentiment classification [18], and the role of large language models in sentiment analysis [16]. These studies collectively highlight the growing importance of hybrid approaches that combine the contextual understanding of Transformer models with the sequential learning capabilities of recurrent networks.

III. PROPOSED METHODOLOGY

This section presents the proposed hybrid deep learning approach for enhanced sentiment analysis. The system architecture integrates RoBERTa for contextual feature extraction with BiLSTM for bidirectional sequence learning, followed by classification through fully connected and softmax layers.

A. System Architecture

The proposed system follows a multi-stage pipeline comprising data collection, data preprocessing, feature extraction using RoBERTa, sequence learning using BiLSTM, and classification. The preprocessed text is first tokenized using the RoBERTa tokenizer and passed through the RoBERTa encoder to generate 768-dimensional contextual embeddings. These embeddings are then fed into a BiLSTM layer that processes the sequence in both forward and backward directions. A dropout layer is applied for regularization, followed by a fully connected layer and a softmax layer for final classification into three sentiment categories: positive, negative, and neutral.

B. Data Collection

The Twitter US Airline Sentiment dataset is used as the primary data source in this study. This publicly available dataset, obtained from Kaggle, contains 14,299 labeled tweets representing sentiments expressed by passengers towards major US airlines. The dataset includes three sentiment classes: negative (comprising the majority of samples), neutral, and positive. The dataset provides real-world textual data including abbreviations, slang, and informal expressions, making it suitable for evaluating sentiment analysis models on social media content.

C. Data Preprocessing

Data preprocessing is performed to improve the quality of input text. Raw tweet data contains noise such as Twitter handles (@username), URLs, hashtags, special characters, and irrelevant symbols that can adversely affect model performance. The preprocessing pipeline includes the following operations: removal of Twitter handles, removal of URLs and links, removal of hashtag symbols while retaining the text, removal of special characters and punctuation, conversion to lowercase, and removal of extra whitespace. Additionally, the cleaned text is tokenized using the RoBERTa tokenizer (Byte-Pair Encoding) with a maximum sequence length of 128 tokens. Padding and truncation are applied to ensure uniform input dimensions. The sentiment labels are encoded numerically, and the dataset is split into training (80%) and testing (20%) sets using stratified sampling to maintain balanced class distribution.

D. RoBERTa Model

RoBERTa (Robustly Optimized BERT Pretraining Approach) is a Transformer-based model used for generating contextual embeddings [1]. It is an improved variant of BERT that employs optimized pretraining strategies including training with larger mini-batches, longer sequences, dynamically changing the masking pattern, and removing the next sentence prediction objective. In this work, the pretrained roberta-base model with approximately 125 million parameters is utilized. The model produces 768-dimensional embeddings that capture the semantic relationships between words based on their surrounding context. RoBERTa processes the entire input sequence in parallel through its self-attention mechanism, making it more computationally efficient than traditional sequential models. During training, the RoBERTa encoder is fine-tuned along with the downstream BiLSTM layers to adapt the contextual representations to the sentiment analysis task.

E. BiLSTM Model

BiLSTM (Bidirectional Long Short-Term Memory) is a type of recurrent neural network designed for sequence modeling [5]. Unlike standard LSTMs that process input in a single direction, BiLSTM processes the input sequence in both forward and backward directions simultaneously, enabling the model to capture dependencies from both past and future context. BiLSTM consists of memory cells with gating mechanisms (input, forget, and output gates) that retain important information over long sequences, effectively addressing the vanishing gradient problem that affects standard RNNs. In the proposed architecture, a two-layer BiLSTM with 256 hidden units per direction is employed after the RoBERTa embedding layer. The forward and backward hidden states are concatenated, producing a total hidden output of 512 dimensions (256 multiplied by 2). A dropout of 0.3 is applied between BiLSTM layers to prevent overfitting.

F. Classification Head

The classification head receives the combined BiLSTM output and passes it through a dropout layer (rate 0.3) followed by a fully connected layer that maps the 512-dimensional features to three output classes. The softmax activation function converts the output logits into probability values for each sentiment class, and the class with the highest probability is selected as the predicted sentiment.

G. *Model Training*

The hybrid model is trained using the processed dataset with the following training configuration: the number of training epochs is set to 5, the batch size is 32, and the learning rate is 2e-5 with a linear warmup scheduler. The AdamW optimizer is used for parameter optimization. CrossEntropyLoss serves as the loss function, and gradient clipping with a maximum norm of 1.0 is applied to stabilize training. The model is trained on GPU hardware when available to accelerate computation. The evaluation metrics used include accuracy, precision, recall, and F1-score.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental setup, evaluation metrics, and results obtained from the proposed RoBERTa-BiLSTM hybrid model.

A. *Dataset Statistics*

The Twitter US Airline Sentiment dataset comprises 14,299 samples distributed across three classes. The training set contains 11,439 samples, and the test set contains 2,860 samples. The training set distribution is as follows: 7,309 negative samples, 2,366 neutral samples, and 1,764 positive samples. The test set comprises 1,828 negative, 591 neutral, and 441 positive samples. The imbalanced nature of the dataset, with negative tweets forming the majority, presents a realistic challenge for classification.

B. *Performance Metrics*

The proposed model is evaluated using standard classification metrics. The overall model performance on the test set is summarized as follows: the model achieves an accuracy of 85.14%, a weighted precision of 0.8478, a weighted recall of 0.8514, and a weighted F1-score of 0.8487. The macro-averaged precision, recall, and F1-score are 0.8158, 0.7854, and 0.7994 respectively, indicating consistent performance across all three classes despite the class imbalance in the dataset.

C. *Per-Class Classification Results*

The per-class classification results reveal the following performance characteristics. For the negative class (1,828 test samples), the model achieves a precision of 0.8914, a recall of 0.9344, and an F1-score of 0.9124. For the neutral class (591 test samples), the precision is 0.7256, the recall is 0.6667, and the F1-score is 0.6949. For the positive class (441 test samples), the precision is 0.8304, the recall is 0.7551, and the F1-score is 0.7910. These results indicate that the model performs best on the negative class due to the larger number of training samples, while the neutral class is the most challenging to classify, which is consistent with observations in the literature [1], [4].

TABLE I
CLASSIFICATION REPORT OF THE PROPOSED MODEL

Class	Precision	Recall	F1-Score	Support
Negative	0.8914	0.9344	0.9124	1828
Neutral	0.7256	0.6667	0.6949	591
Positive	0.8304	0.7551	0.7910	441
Accuracy			0.8514	2860
Macro Avg	0.8158	0.7854	0.7994	2860
Weighted Avg	0.8478	0.8514	0.8487	2860

D. *Training History Analysis*

The training history over 5 epochs reveals the convergence behavior of the model. The training loss decreases consistently across epochs, indicating effective learning. The validation loss shows a decreasing trend in early epochs but exhibits slight fluctuation in later epochs, which is typical of fine-tuning pretrained models. The training accuracy increases progressively, while the validation accuracy reaches its best value of 0.8531 during training. The gap between training and validation accuracy remains relatively small, suggesting that the dropout regularization and gradient clipping effectively mitigate overfitting.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

E. Discussion

The experimental results demonstrate that the proposed RoBERTa-BiLSTM hybrid model provides effective sentiment classification on the Twitter US Airline Sentiment dataset. The integration of RoBERTa for contextual understanding with BiLSTM for sequential dependency capture proves beneficial compared to using either component independently. RoBERTa provides rich contextual embeddings through its self-attention mechanism and pretrained knowledge, while BiLSTM enhances the model by capturing forward and backward sequential relationships between words.

The model performs best on the negative class (F1-score 0.9124), which benefits from the larger proportion of negative samples in the training set. The positive class achieves a reasonable F1-score of 0.7910, while the neutral class presents the most challenge with an F1-score of 0.6949. This difficulty in classifying neutral sentiments is well-documented in the literature, as neutral text often shares linguistic features with both positive and negative categories [4], [8]. The overall accuracy of 85.14% and weighted F1-score of 0.8487 demonstrate the practical applicability of the proposed approach for real-world sentiment analysis tasks.

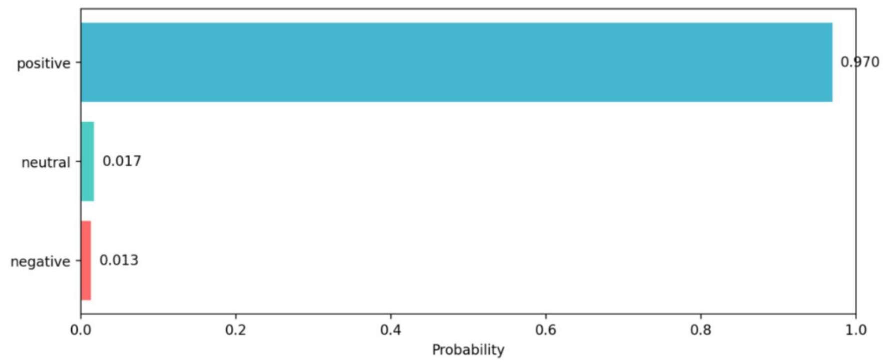

Figure 1. shows the probability distribution of the sample text prediction

The system was implemented using Python with PyTorch as the deep learning framework and Hugging Face Transformers for the pretrained RoBERTa model. A Streamlit-based web application was developed to provide an interactive interface supporting both single text prediction and batch prediction functionality, enabling users to analyze sentiments with confidence scores and visual representations of results.

V. CONCLUSION AND FUTURE WORK

This paper presented an enhanced sentiment analysis system using a hybrid deep learning approach combining RoBERTa and BiLSTM. The proposed model leverages the contextual understanding capability of RoBERTa through its Transformer-based self-attention mechanism and the sequential learning capability of BiLSTM to capture bidirectional dependencies in text. The model was trained and evaluated on the Twitter US Airline Sentiment dataset containing 14,299 samples across three sentiment categories.

The experimental results demonstrate that the proposed hybrid model achieves an accuracy of 85.14%, a weighted precision of 0.8478, a weighted recall of 0.8514, and a weighted F1-score of 0.8487. The per-class analysis shows strong performance on the negative class (F1-score of 0.9124) with competitive results on the positive class (F1-score of 0.7910). The system was implemented as a Streamlit-based web application supporting both single text and batch prediction with confidence scores and visualizations.

Future work includes extending the system to support multilingual sentiment analysis for processing text in multiple languages. The model can be enhanced by integrating real-time data from social media platforms for live sentiment monitoring. Deployment as a full-scale web or mobile application can improve accessibility and usability. Furthermore, incorporating attention mechanisms and explainable AI methods can provide better interpretability of predictions. Domain-specific sentiment analysis for healthcare, finance, and other specialized domains can also be explored. Advanced ensemble techniques and larger pretrained models may further improve classification performance.

VI. ACKNOWLEDGMENT

The authors wish to acknowledge the management and faculty of Geethanjali Institute of Science and Technology, Nellore, for providing the necessary facilities and support for this research work.

REFERENCES

[1] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," IEEE Trans. Emerg. Topics Comput. Intell., vol. 9, no. 6, pp. 3788–3799, Dec. 2025.

[2] Z. Sokolová, M. Harahus, M. Sokol, E. Kupcová, and M. Pleva, "Sentiment analysis using transformer models: BERT, T5, and GPT," in Proc. 35th Int. Conf. Radioelektronika (RADIOELEKTRONIKA), IEEE, 2025.

[3] X. Zhang, Y. Liu, T. Zhang, L. Hou, X. Liu, Z. Guo, and A. Mulati, "A BERT–LSTM–Attention framework for robust multi-class sentiment analysis on Twitter data," Systems, vol. 13, no. 11, p. 964, 2025.

[4] S. B. Bhonde and J. R. Prasad, "Sentiment analysis—Methods, applications & challenges," Int. J. Electron. Commun. Comput. Eng., vol. 6, no. 6, pp. 634–640, 2015.

[5] R. T. Younesi, J. Tanha, S. Namvar, and S. H. Mostafaei, "A CNN-BiLSTM based deep learning model for sentiment analysis," in Proc. 20th CSI Int. Symp. Artif. Intell. Signal Process., 2023.

[6] R. Pandey and J. P. Singh, "BERT-LSTM model for sarcasm detection in code-mixed social media post," J. Intell. Inf. Syst., vol. 60, no. 1, pp. 235–254, 2023.

[7] S. Tam, R. B. Said, and Ö. Ö. Tanriöver, "A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification," IEEE Access, vol. 9, pp. 41283–41293, 2021.

[8] Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa, and J. Baldwin, "Sentiment analysis of comments in social media," Int. J. Elect. Comput. Eng., vol. 10, no. 6, pp. 2088–8708, 2020.

[9] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 6000–6010.

[10] I. Saad, "Opinion mining on US airline Twitter data using machine learning techniques," in Proc. 16th Int. Comput. Eng. Conf., 2020, pp. 59–63.

[11] E. Prabhakar, M. Santhosh, A. H. Krishnan, T. Kumar, and R. Sudhakar, "Sentiment analysis of US airline Twitter data using new AdaBoost approach," Int. J. Eng. Res. Technol., vol. 7, no. 1, pp. 1–6, 2019.

[12] K. Madhuri, "A machine learning based framework for sentiment classification: Indian railways case study," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 4, pp. 441–445, 2019.

[13] M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM algorithm based on sentiment analysis using review dataset," in Proc. 8th Int. Conf. System Model. Advance. Res. Trends, 2019, pp. 266–270.

[14] R. M. Alahmary, H. Z. Al-Dossari, and A. Z. Emam, "Sentiment analysis of Saudi dialect using deep learning techniques," in Proc. Int. Conf. Electron., Inf., Commun., 2019, pp. 1–6.

[15] T. Dholpuria, Y. Rana, and C. Agrawal, "A sentiment analysis approach through deep learning for a movie review," in Proc. 8th Int. Conf. Commun. Syst. Netw. Technol., 2018, pp. 173–181.

[16] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in Findings Assoc. Comput. Linguistics: NAACL 2024, 2024.

[17] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods," IEEE Trans. Comput. Social Syst., vol. 7, no. 6, pp. 1358–1375, Dec. 2020.

[18] A. Mabrouk, R. P. D. Redondo, and M. Kayed, "Deep learning-based sentiment classification: A comparative survey," IEEE Access, vol. 8, pp. 85616–85638, 2020.

[19] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," Appl. Intell., vol. 51, pp. 3522–3533, 2021.

[20] P. Bhuvaneshwari, A. N. Rao, Y. H. Robinson, and M. Thippeswamy, "Sentiment analysis for user reviews using Bi-LSTM self-attention based CNN model," Multimedia Tools Appl., vol. 81, no. 9, pp. 12405–12419, 2022.