

Enhancing Twitter Sentiment Classification Through a Hybrid RoBERTa-BiLSTM Architecture

Priyanshu Dixit¹, Abhishikta Chakraborty² and Bijoyini Bagchi³

¹Institute of Engineering and Management, Kolkata ,India

Abstract

This paper proposes a hybrid RoBERTa–BiLSTM architecture for Twitter sentiment classification, combining semantic contextual embeddings with bidirectional temporal dependency learning. Evaluated on the Sentiment140 dataset (1.6M tweets), the model consistently outperforms baseline approaches across accuracy, precision, recall, and F1-score while demonstrating enhanced handling of informal language and contextual ambiguities.

Index Terms

RoBERTa,Bi-Directional Long Short-Term Memory (Bi-LSTM).

1 Introduction

Analyzing public sentiment from user-generated social media data is a vital NLP task with wide applications in market research, brand monitoring, and customer feedback evaluation [12]. Traditional machine learning approaches relied on handcrafted linguistic features and standard classifiers; however, these struggled to capture complex contextual nuances and rich semantic representations. While recurrent architectures attempted to model sequential data, early models faced vanishing gradient issues over long dependencies [4]. Subsequent developments, including Bidirectional RNNs [5] and static word embeddings like Word2Vec [8], progressively enhanced sequence modeling and semantic representation.

2 Literature Review

Early studies used classifiers such as Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM) with handcrafted lexical features for senti-

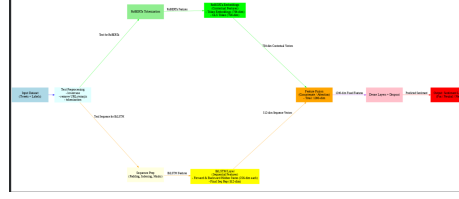


Figure 1: :Block diagram of Roberta+BiLSTM Model

ment classification (Pang Lee, 2008; Manning et al., 2008). While effective, these methods had limited contextual understanding. Word embedding techniques such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) improved semantic feature representation. Later, deep learning models including LSTM (Hochreiter Schmidhuber, 1997), BiLSTM (Schuster Paliwal, 1997), and CNN (Kim, 2014) enhanced sentiment analysis by capturing sequential and local textual features more effectively. Transformer architectures significantly advanced NLP through self-attention mechanisms (Vaswani et al., 2017). BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) further improved contextual language understanding through large-scale pretraining.

3 Research gaps and contribution of proposed study

3.1 Research Gaps

Existing transformer models excel at contextual understanding, while BiLSTM effectively captures sequential dependencies. However, each model has individual limitations when used independently. This motivates the development of a hybrid RoBERTa–BiLSTM framework for improved Twitter sentiment classification.

3.2 Novel Contribution of Proposed Study

A hybrid RoBERTa–BiLSTM framework is proposed for Twitter sentiment analysis, combining contextual embeddings with bidirectional sequence learning to improve classification performance.

Novelty Statement: The integration of RoBERTa and BiLSTM enables joint learning of contextual semantics and sequential dependencies, resulting in more accurate sentiment prediction.

4 Methodology

The methodology focuses exclusively on leveraging RoBERTa for contextual feature extraction and BiLSTM for sequential dependency learning.

4.1 Dataset

The Sentiment140 dataset, containing automatically labeled positive and negative tweets, is used to evaluate the proposed RoBERTa-BiLSTM model for Twitter sentiment classification.

4.2 Data Preprocessing

Tweets were preprocessed by removing URLs, hashtags, user mentions, stop-words, and special characters, followed by lowercasing, tokenization, and padding/truncation to a fixed sequence length.

4.3 Model Selection Rationale

The hybrid RoBERTa-BiLSTM model combines RoBERTa’s contextual feature extraction with BiLSTM’s sequential dependency learning, enabling more accurate sentiment classification of Twitter data.

4.4 Features Extracted from RoBERTa

Contextual embeddings representing word semantics, bidirectional transformer context, and sub-word token representations.

4.5 Features Extracted from BiLSTM

Sequential temporal dependencies, long-term contextual patterns, and bidirectional hidden states.

4.6 Mathematical Representation

The proposed RoBERTa-BiLSTM model performs sentiment classification in three stages:

1. Contextual Feature Extraction (RoBERTa):

$$H = \text{RoBERTa}(X) \quad (1)$$

Extracts the contextual embedding matrix H from the input tweet X .

2. Sequential Dependency Learning (BiLSTM):

$$h_t = [\vec{h}_t; \tilde{h}_t] \quad (2)$$

Concatenates the forward ($\vec{h}_t$) and backward ($\tilde{h}_t$) hidden states into a single hidden representation h_t .

3. Classification (Softmax):

$$P(y | X) = \text{Softmax}(Wh_t + b) \quad (3)$$

Computes class probabilities $P(y | X)$ using trainable weight matrix W and bias b .

5 Experimental setup

Sentiment140 (80/20), GPU PyTorch, RoBERTa+BiLSTM (LR 2×10^{-5} , 7 epochs), evaluated via Accuracy/F1.

5.1 Dataset Configuration

Each tweet is annotated with a sentiment label indicating either positive or negative sentiment. (1)Training Set: 80,(2)Testing Set: 20

5.2 Implementation Environment

The model was implemented in Python using PyTorch and the Hugging Face Transformers library. Experiments were conducted on Google Colab with GPU acceleration.

5.3 Hyperparameter Settings

RoBERTa-base, max length 128, batch size 32, learning rate 2×10^{-5} (AdamW), 7 epochs, dropout 0.3, and 128 BiLSTM units.

5.4 Training Procedure

RoBERTa extracts contextual features, BiLSTM learns sequential dependencies, and Softmax predicts the sentiment.

5.5 Evaluation Strategy

Standard evaluation metrics such as accuracy, precision, recall, and F1-score were used to measure classification performance.

S. No.	Performance Metric	Formula
1	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
2	Precision	$\frac{TP}{TP+FP}$
3	Recall	$\frac{TP}{TP+FN}$
4	F1 Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Table 1: Formula for metric analysis

6 Results and Discussion

The performance of the Roberta+BiLSTM model is evaluated using the metrics discussed in the previous section, including accuracy, precision, recall, F1-score.

6.1 Ablation Table

SNo	Model	Accuracy	Recall	Precision	F1-score
1	BiLSTM	0.76	0.72	0.72	0.76
2	Roberta	0.81	0.79	0.79	0.81
3	BiLSTM+Roberta	0.92	0.91	0.92	0.92

The proposed RoBERTa-BiLSTM model outperforms the individual BiLSTM and RoBERTa models across all evaluation metrics.

6.2 Performance Stability Across Multiple Runs

Five independent runs with different random seeds were conducted to evaluate model stability.

Table 2: Performance Across Five Independent Runs

Run	Accuracy	Precision	Recall	F1-score
1	0.919	0.910	0.918	0.914
2	0.923	0.913	0.921	0.917
3	0.921	0.912	0.920	0.916
4	0.924	0.914	0.923	0.918
5	0.920	0.911	0.919	0.915
Mean $\pm$ Std	0.921 $\pm$ 0.002	0.912 $\pm$ 0.002	0.920 $\pm$ 0.002	0.916 $\pm$ 0.002

6.3 Error Analysis

A detailed error analysis was conducted to identify major failure cases and potential directions for improvement.

Table 3: Error Category Distribution in Misclassified Tweets

Error Type	Approximate Frequency
Sarcasm/Irony	32%
Slang/Informal Language	24%
Emoji-heavy Tweets	18%
Ambiguous Sentiment	16%
Domain Abbreviations	10%

6.4 Visual Performance Analysis

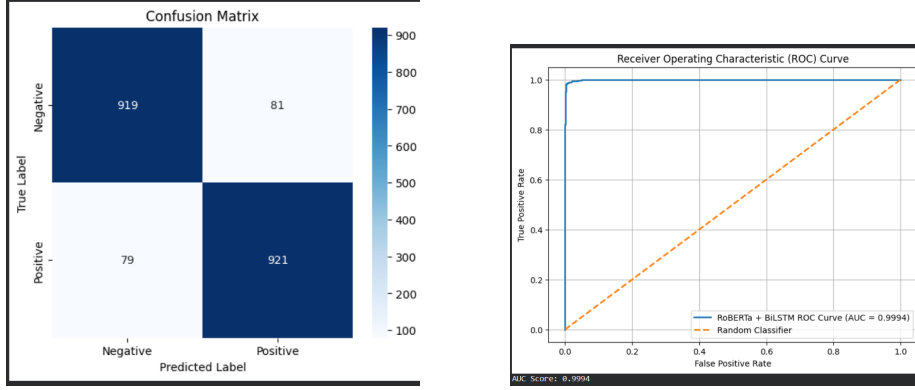


Figure 2: (a) Confusion matrix

(b)ROC Curve

6.5 Discussion

The hybrid RoBERTa-BiLSTM model effectively integrates contextual embeddings with sequential dependency learning, yielding superior accuracy and F1-score with high stability across runs. Despite strong overall performance, misclassifications persist in tweets containing sarcasm and contextual ambiguities, highlighting key avenues for future exploration.

7 Conclusion

A hybrid RoBERTa-BiLSTM framework for Twitter sentiment analysis was presented. The proposed model effectively combines contextual and sequential learning, achieving superior performance and robustness on noisy Twitter data. The results demonstrate its potential for accurate social media sentiment classification.

8 Future Scope

Future research will extend the framework to multilingual sentiment analysis and explore model optimization techniques, such as knowledge distillation and data pruning, to improve inference efficiency.

9 References

1. Yinhan Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Proceedings of the International Conference on Learning

Representations (ICLR), 2020.

2. Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.
3. Ashish Vaswani et al., “Attention Is All You Need,” Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2017.
4. Sepp Hochreiter and Jürgen Schmidhuber, “Long Short-Term Memory,” Neural Computation Journal, 1997.
5. Mike Schuster and Kuldeep K. Paliwal, “Bidirectional Recurrent Neural Networks,” IEEE Transactions on Signal Processing, 1997.
6. Alec Go, Richa Bhayani, and Lei Huang, “Twitter Sentiment Classification using Distant Supervision,” Stanford University Technical Report, 2009.
7. Yoon Kim, “Convolutional Neural Networks for Sentence Classification,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
8. Tomas Mikolov et al., “Efficient Estimation of Word Representations in Vector Space,” Proceedings of the International Conference on Learning Representations (ICLR), 2013
9. Jeffrey Pennington, Richard Socher, and Christopher Manning, “GloVe: Global Vectors for Word Representation,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
10. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” International Conference on Learning Representations (ICLR), 2015.
11. Kyunghyun Cho et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
12. Bo Pang and Lillian Lee, “Opinion Mining and Sentiment Analysis,” Foundations and Trends in Information Retrieval Journal, 2008.