IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15422

Sentiment Analysis Using RoBERTa-Based Hybrid Model

M. Srivalli[1], A. Sri Nagesh[2], Jaideep Gera[3], Sk. Moinuddin[4], P. Gnana Suji[5]

B.Tech Students, Department of CSBS, R.V.R & J.C. College of Engineering, Guntur, India[1,4,5]

Professor & Head of Department, Department of CSBS, R.V.R & J.C. College of Engineering, Guntur, India[2]

Associate Professor, Department of CSBS, R.V.R & J.C. College of Engineering, Guntur, India[3]

Abstract: The sentiment analysis of movie reviews is a well-known issue within the Natural Language Processing (NLP) field and finds the primary applications in opinion mining and movie recommendation systems. Over the last few years, a number of studies have investigated the concept of hybrid deep learning structures that integrate transformer encoders with recurrent and structured prediction stack to enhance sentiment classification. The transformer encoder in the given study is substituted with a Robustly Optimized BERT pretraining approach (RoBERTa) and LSTM with a Bidirectional Long Short-Term Memory (BiLSTM) network. The suggested method combines RoBERTa-BiLSTM and a Conditional Random Field (CRF) layer to maintain consistency with base architecture. The suggested framework is tested on the IMDb movie review data set with a structured train validation test splits. The evaluation of performance is done by applying standard classification measures. The model based on RoBERTa achieves an accuracy of 91.01, a precision of 91.43, a recall of 90.50 and an F1-score of 90.97, and the results were higher than the results reported on the Transformer-LSTM-CRF. These results imply that improved contextual representations supplied by the contemporary pre-trained transformers have a positive effect on document-level sentiment classification.

Keywords: Sentiment Analysis, Deep Learning (DL), RoBERTa, BiLSTM, CRF, IMDb Dataset.

I. INTRODUCTION

Sentiment analysis is a field of research that seeks to automatically detect expressions of opinions and emotions in a text where it is commonly used in Natural Language Processing (NLP) because of the popularity of online platforms like movie review websites, e-commerce portals and social media [1]. Long and elaborate opinion statements are common in movie reviews, and thus document level sentiment classification is a daunting task. Past methods of sentiment analysis were primarily rule based and lexicon based methods, which were simple to use, yet they were not able to contextualize meaning and language usage variations [1]. Conventional machine learning techniques made things better by learning with labeled data, however, they needed manual feature engineering and was not as useful when working with long and complex texts. In the recent advances in DL, the ability to directly learn meaningful representations directly using raw text has greatly enhanced sentiment classification. Transformer based models, especially, have demonstrated a high aptitude to communicate contextual data throughout entire texts employing self-attention mechanisms [2]. Nevertheless, the combination of such representations to hybrid classification systems is also a field that should be further investigated empirically. Inspired by the previous work of transformer-LSTM-CRF based sentiment analysis [3], this paper aims at assessing whether the document level sentiment classification can be further enhanced by substituting the transformer encoder with a powerful pre-trained model like RoBERTa on IMDB movie review dataset.

II. RELATED WORK

Sentiment analysis has been extensively studied because of its relevance in the interpretation of opinions that are stated within text data. The initial methods were based on lexicon-based and rule-driven algorithms in which associated sentiment dictionaries were consulted to find out the polarity [1]. Although these techniques were readable and easy to apply, they had the difficulty of grasping contextual nuances and domain specific expressions. The introduction of labeled datasets led to the implementation of more traditional machine learning based classifiers like the Naive Bayes and Support Vector Machines which showed improvements and necessary states of hand crafted feature engineering became common [1].

The introduction of DL models made a substantial contribution to sentiment classification as it helped to extract features in raw text automatically. Convolutional Neural Networks (CNN) has been shown to be effective in the local semantic patterns capture [4], whereas Recurrent Neural Networks and Long Short-Term memory (LSTM) networks were able to

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15422

model sequential information related to context without a problem [5]. Nonetheless, recurrent architectures were limited in the processing of long documents because of the problems of modeling of long range dependencies. Transformer-based designs have overcome these difficulties by using self attention mechanisms that effectively encode global context in sequences [2]. Later research emphasized transformers as the leading models to perform several natural language processing tasks [6], and pre-trained models like BERT performed even better with the help of large-scale training on unlabeled corpora [7].

RoBERTa, which built upon these advancements, proposed optimized pre-training methods and showed good results in document-level sentiment classification tasks, even on benchmark datasets, such as IMDb [8], [9]. In structured prediction tasks, Conditional Random Fields (CRFs) are usually combined to produce dependence among labels on the output [10]. Recent transformer-LSTM-CRF architecture achieved better scores on various datasets [3]. Nevertheless, the role of substituting the previous transformer encoders with the stronger pre-trained models such as RoBERTa in this hybrid architecture is less studied, which explains the current empirical study.

III. METHODOLOGY

The segment outlines the suggested sentiment analysis framework, its general workflow, preprocessing of the data, as well as the model structure with training details utilized to conduct the review sentiment analysis.

A. Proposed Framework

The current solution is based on a hybrid deep learning model that is aimed at classifying sentiments. The primary aim is to determine how the transformer encoder may be substituted with RoBERTa in a transformer-LSTM-CRF framework and the rest of the elements are held constant to compare the results with each other.

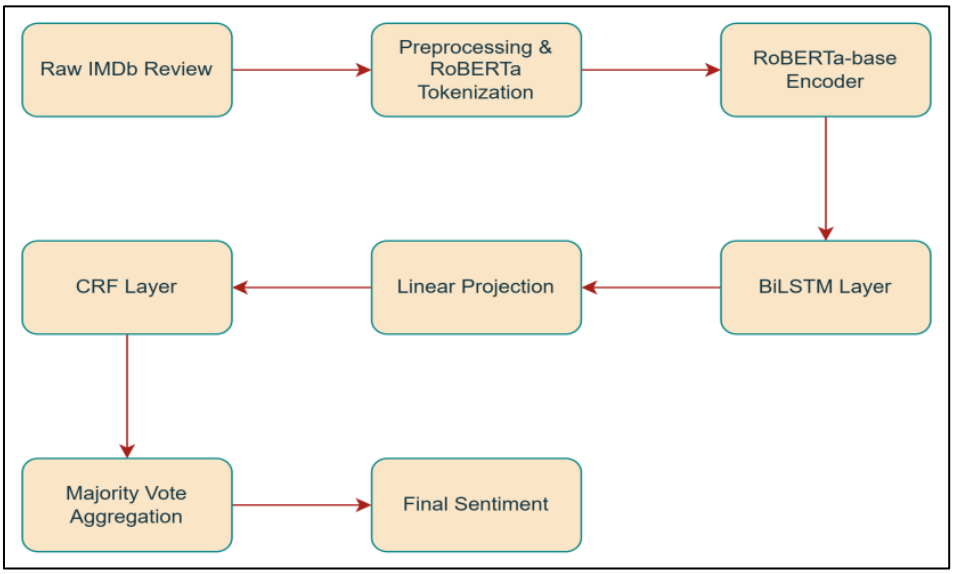

Fig. 1 Flow diagram of RoBERTa-BiLSTM-CRF based sentiment analysis model

The workflow shown in Fig. 1 starts with text preprocessing and tokenization and then goes on to extract contextual features with RoBERTa. The obtained representations are then optimized with sequence modeling based on a BiLSTM network and subsequent sentiment prediction on the basis of a Conditional Random Field (CRF) layer. This hierarchical workflow facilitates the modeling of situational and chronological information that are found in movie reviews.

B. Preprocessing and Tokenization of Data

In the IMDb dataset moview review is considered as a single document. To remove noise without losing the actual meaning, minimal preprocessing is done. This involves the transformation of text into lowcase, the loss of HTML tags and URLs and the regularization of large spacing between words. There is no aggressive linguistic preprocessing, e.g. stemming or lemmatization. The text is then tokenized with the RoBERTa tokenizer that has subword level tokenization to successfully deal with out-of-vocabulary wording. The model needs some special tokens and they are automatically

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15422

added and padding/ truncation forces all the sequences to the same fixed length of 128 tokens to have the same representation of the input in training.

C. Model Architecture

The main idea of our system is core feature extraction based on the pre-trained RoBERTa model. This encoder produces rich semantic meaning of the entire review by giving dense hidden state vectors when fed with the padded 128-token sequences. The embeddings are then passed into a BiLSTM network. Considering both the left-to-right and right-to-left evaluation of the text at the same time, this repeated element is useful to visually chart the chronological movement and long-term relationships of the opinion of the user. Lastly, as in the construction of the baseline study, we use a CRF layer to make the final classification. The raw emission logits of the recurrent layer enters the CRF layer which imposes structural constraints to find the most likely sequence of sentiment tags. In order to generalize this token level output to document-level analysis, we combine the sequence of predicted outputs through a majority voting scheme. This architecture is formalized in Algorithm 1, which shows all the data flow of the architecture.

Algorithm 1 Sentiment Classification using RoBERTa-BiLSTM-CRF

Require: reviewText
Ensure: sentimentLabel
1: tokens := RobertaTokenizer(reviewText, maxLen=128)
2: contextFeatures := RoBERTaEncoder(tokens)
3: lstmOutput := BiLSTM(contextFeatures)
4: **if** phase == Training **then**
5: crfLoss :=NegativeLogLikelihood(lstmOutput, trueLabel)
6: OptimizeParameters(crfLoss)
7: **else if** phase == Evaluation **then**
8: tagSequence := ViterbiAlgorithm(lstmOutput)
9: sentimentLabel:= CalculateMajorityVote(tagSequence)
10: **return** sentimentLabel

IV. EXPERIMENTAL SETUP AND RESULTS

Our test on this architecture was done on IMDb dataset. It is important to note that although the baseline study used a training to testing ratio of 80:20, our experiment used a much more stringent 50:50 split. Out of the entire corpus, 25000 reviews were precisely isolated to be tested. Out of the 25000 training samples, 10% were reserved as a validation subset to monitor the epoch level convergence and avoid overfitting. The test set was not shown at all until the last performance evaluation.

PyTorch is used to perform model implementation and experimentation and transformer components are supported by the Hugging Face library. RoBERTa-base is used as the pre-trained transformer encoder. We are counting on optimizing the model to a number of five epochs with AdamW (learn rate 2 x 10^{-5}), batch size 16, and a 128-token constraint to decrease negative log-likelihood loss of the CRF layer.

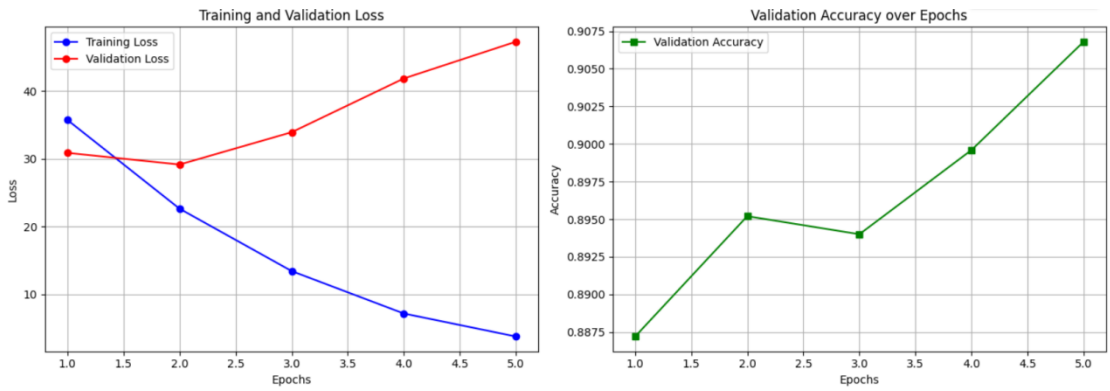

Fig. 2 Training,validation Loss vs. Epochs and validations Accuracy vs. Epochs

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15422

In Fig. 2, during training, a steady reduction in training loss over training epochs is observed in the model, which implies that model parameters are optimized. The validation accuracy increases with the successive epochs, and the behavior of generalization remains steady as well. Whereas the loss of validation is not exactly the same trend as validation accuracy in later epochs, this behavior should be seen in a document level classification scenario using a CRF layer, where loss is calculated at the sequence level, and evaluation at the document level.

Once training is over, the end model is evaluated using the test dataset. The developed RoBERTa-BiLSTM-CRF model provides a general classification performance of about 91.01 which is quite satisfactory in terms of document level sentiment classification.

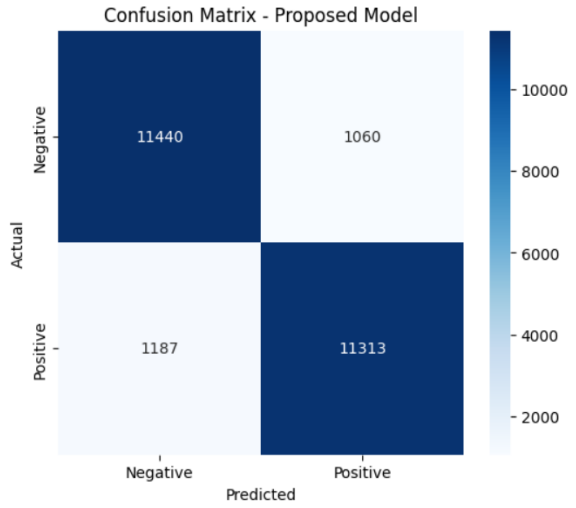

Fig. 3 Roberta-BiLSTM-CRF model confusion matrix

The confusion matrix in Fig. 3 indicates that the model accurately predicts most positive and negative reviews and the misclassification rates of the classes are rather equal. This implies that there is no tremendous bias of the proposed model in relation to either category of sentiment.

In order to further evaluate effectiveness, performance of current model is compared to that of base transformer-LSTM-CRF model in the previous work. Base model results are not reimplemented in this work, rather taken directly out of the respective study. As shown in Table I, substituting the initial transformer encoder with RoBERTa results in an evident performance improvement in all evaluation metrics.

Table I Performance to Base Model

Metric	Transformer-LSTM-CRF [3]	RoBERTa-BiLSTM-CRF (Proposed)	Improvement
Accuracy %	85.51	91.01	5.50
Precision %	85.51	91.43	5.92
Recall %	83.77	90.50	6.73
F1-Score %	85.06	90.97	5.91

V. CONCLUSION

This paper had provided an empirical investigation of a RoBERTa-based hybrid framework of document-level sentiment recognition with the use of the IMDb dataset. This study tested the effectiveness of the modern pre-trained language models using complex structured prediction assignments by modifying an existing Transformer-LSTM-CRF set-up and replacing the original encoder with RoBERTa. As it has been proved through the results of the experiments, the proposed RoBERTa-BiLSTM-CRF model does not only reach a much higher accuracy (at 91.01 percent) than the baseline

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15422

Transformer-LSTM-CRF model, but also requires significantly less training data. These results confirm that rich contextual representations produced by RoBERTa influence the improvement in performance in processing long and intricate reviews, whereas the BiLSTM and CRF components effectively provide sequential and structural uniformity. Although the current research concentrates on document-level classification, future studies ought to be done on sentiment analysis at a sentence or aspect level at which CRFs can be used to provide even more benefits. Also, it will be possible to test the framework on benchmark datasets across various domains to identify its generalization abilities. Lastly, it would be interesting to investigate larger variants of transformers, other pooling methods or even computational optimization methods to get even greater performance and efficiency and deploy it in the real world.

REFERENCES

[1]. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
[2]. A. Vaswani et al., "Attention is all you need," *in Proceedings of the Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA*, 2017, pp. 5998–6008.
[3]. L. Yao and N. Zheng, "Sentiment Analysis Based on Improved Transformer Model and Conditional Random Fields," *IEEE Access*, vol. 12, pp. 90145–90156, 2024.
[4]. Y. Kim, "Convolutional neural networks for sentence classification," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar*, 2014, pp. 1746–1751.
[5]. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[6]. T. Wolf et al., "Transformers: State-of-the-art natural language processing," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2020, pp. 38–45.
[7]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *in Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA*, 2019, pp. 4171–4186.
[8]. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[9]. A. L. Maas et al., "Learning word vectors for sentiment analysis," *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Portland, OR, USA*, 2011, pp. 142–150.
[10]. J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *in Proceedings of the International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

