Implementation of BERT-BiLSTM Model for Sentiment Classification of Movie Reviews on Rotten Tomatoes

Thoriq Gifari Nurizky[1], Endah Setyowati[2*]

[1,2]Telecommunication System Department, Universitas Pendidikan Indonesia, Bandung, West Java, Indonesia
Email: [1]thorizky@upi.edu, [2*]endahsetyowati@upi.edu

(Received: 4 Dec 2025, revised: 2 Mar 2026, 18 Mar 2026, accepted: 26 Mar 2026)

Abstract

This study aims to analyze movie review sentiment using a Bidirectional Encoder Representations from Transformers-Bidirectional Long Short-Term Memory (BERT-BiLSTM) hybrid model designed to capture global contextual information and local sequential dependencies in complex texts. The dataset is obtained from Rotten Tomatoes Critic Reviews, consisting of 1,048,576 entries, of which 10,000 samples are randomly selected due to computational limitations. The research methodology includes text preprocessing, tokenization with the BERT tokenizer, model training, and performance evaluation using accuracy, precision, recall, and F1-score. The implementation uses the AdamW optimizer with a learning rate of 2×10^{-5} for three epochs on an AMD Ryzen 3 5300U CPU. For comparison, the hybrid model is evaluated alongside a single BERT model and a single BiLSTM model. Experimental results show that the BERT-BiLSTM model achieves the highest performance, obtaining an accuracy, precision, recall, and F1-score of 84%. In contrast, the single BERT model achieved 66% accuracy and an F1 score of 57%, while the single BiLSTM model achieved 63% accuracy and an F1 score of 48%. These results indicate that combining transformer-based contextual representation with bidirectional sequence modeling improves the effectiveness of sentiment classification compared to the single-model approach. Therefore, the BERT-BiLSTM hybrid method proves effective for analyzing complex textual sentiment and has potential applications across various review domains that require deep contextual understanding.

Keywords: BERT, BiLSTM, Deep Learning, Movie Reviews, Sentiment Analysis

I. INTRODUCTION

The rapid growth of digital platforms offering such review features has generated vast collections of public opinion, a potentially valuable source of information for business decision-makers, policymakers, and researchers. Sentiment analysis aims to extract opinion polarity (such as positive or negative) from unstructured text, thereby providing quantitative insights into public perception [1]. Early techniques in sentiment analysis often relied on dictionary-based approaches or traditional machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, which use representations like bag-of-words or TF-IDF. While lightweight, these methods struggle to capture semantic context and long-range relationships among tokens in complex sentences [2].

Advances in deep learning architectures and pre-trained language models have revolutionized natural language processing. Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) introduce contextual and dynamic word representations, where the meaning of a token is bidirectionally influenced by its sentence context. This has been shown to improve performance for various Natural Language Processing (NLP) tasks, including text classification and sentiment analysis. Furthermore, systematic reviews and comparative experiments show that BERT and its variants (such as RoBERTa, DistilBERT, and others) consistently outperform static embeddings (Word2Vec/GloVe/FastText) in sentiment classification tasks on diverse datasets (IMDB, Yelp, Twitter) [3], [4].

Furthermore, sequential models such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) remain important due to their ability to capture text sequence dependencies and long-term contextual patterns. BiLSTM processes token sequences both forward and backward, making them more sensitive to local and global context than unidirectional LSTMs [5], [6]. Several studies combine the advantages of transformers (as encoders or feature extractors) and sequential networks (as sequential processing layers or classifiers) to improve classification performance; The combination of BERT+BiLSTM has been reported to extract rich semantic features from BERT while leveraging its

sequential LSTM mechanism for further processing, often resulting in improved parameter evaluation on diverse datasets [7], [8].

Several related studies have applied the BERT-BiLSTM variant to specific domains. For example, studies on Chinese text and social media have shown that the BERT & BiLSTM workflow (with BERT as the embedding generator and BiLSTM as the sequential feature extractor) improves sentiment classification accuracy compared to either BERT or BiLSTM alone [9], [10]. Similarly, recent studies in the Journal of Big Data and other conference papers present experimental results supporting the idea of improved F1 scores and classification stability on noisy or short text data [11], [12].

Other studies analyzing sentiment using deep learning have shown significant progress in recent years. Research shows that transformer-based models like BERT can improve sentiment accuracy by over 90% in product reviews, primarily due to their bidirectional contextual capabilities [13]. Other relevant research shows that combining BERT with sequential architectures such as LSTM can detect word ambiguity and improve F1 scores on noisy social media data [14]. Further exploring the context of movie reviews, recent research shows that movie review datasets are linguistically more diverse compared to product reviews, suggesting that models combining contextual representation with sequential processing are more effective [15]. Similar findings were also presented by Chen et al. (2021), who explained that the BERT + BiLSTM hybrid model can improve classification accuracy by 4-7% compared to a single model on the Rotten Tomatoes dataset. This study certainly strengthens the relevance of the hybrid approach between BERT and BiLSTM for sentiment analysis in movie reviews, particularly in terms of handling long texts, writing styles, and polarity ambiguities that frequently appear in movie reviews. Therefore, this study adopts this approach as a basis for developing the model for this study.

In the context of movie reviews, datasets like Rotten Tomatoes provide reviews from both critics and the general public. The characteristics of this data, such as diverse writing styles, use of idioms, varying text lengths, and polarity ambiguity, make sentiment classification more challenging than more uniform datasets [16], [17]. Therefore, approaches that combine high-level contextual understanding via BERT and sequence modeling via BiLSTM are relevant and worth testing in this domain. Moreover, unlike static embeddings (such as FastText) that map a single word to a fixed vector, BERT produces representations that vary based on context, which is important for addressing polysemy (words with multiple meanings) and linguistic nuances in movie reviews [18], [19], [20].

Many surveys and benchmark studies suggest strategies or ensembles to improve model robustness on diverse sentiment datasets, including suggestions to experiment with hyperparameters, token length (maximum_length), batch size, and sampling strategy if the original dataset is very large [8], [21]. Furthermore, practical aspects such as resource limitations (GPU/CPU) and dataset size often influence experimental design; some studies suggest using representative samples or lightweight models for proof-of-concept testing before full-scale implementation [22], [23], [24].

Several previous studies have shown that transformer-based models like BERT are capable of generating robust contextual representations in various natural language processing tasks, including sentiment analysis. However, using BERT as a single model still has limitations in capturing complex sequential relationships, especially in long texts. In contrast, recurrent neural network-based models like LSTM or BiLSTM are known to be effective in modeling word-order dependencies, but their performance is highly dependent on the quality of the embedding representation used. Furthermore, many previous studies were conducted with the support of large computing resources, such as the use of GPUs and very large datasets, thus not fully representing the implementation conditions on devices with computing limitations. Research specifically examining the integration of BERT and BiLSTM on the Rotten Tomatoes movie review dataset in a CPU-based computing environment is also relatively limited. Therefore, this study aims to evaluate the performance of a BERT-BiLSTM hybrid model in sentiment analysis of movie reviews using a subset of datasets and a limited computing environment.

Therefore, there remains a research gap in this area. A limited number of studies examine the effectiveness of BERT-BiLSTM hybrid models for sentiment analysis of movie reviews under resource-constrained conditions. This research contributes by implementing and evaluating a BERT-BiLSTM model using samples from the Rotten Tomatoes dataset on devices without GPU support, thus providing added value in terms of both computational efficiency and classification performance.

This study attempts to analyze the performance of a BERT-BiLSTM model in sentiment analysis and movie review classification using the Rotten Tomatoes dataset. This study also examines the extent to which the combination of BERT's contextual embedding and BiLSTM bidirectional sequential modeling contributes to improved evaluation results, particularly in terms of accuracy, precision, recall, and F1-score. Furthermore, this study assesses the model's effectiveness when trained using a limited dataset of 10,000 samples on devices with minimal computing capabilities.

This study aims to design and implement a BERT-BiLSTM hybrid model for sentiment analysis on movie text reviews and assess its performance based on accuracy, precision, recall, and F1 score metrics using the Rotten Tomatoes dataset. Furthermore, this study examines the feasibility of using the model in a computationally constrained environment without dedicated GPU support. Thus, this research is expected to contribute not only to improving sentiment analysis performance but also to the practical aspects of implementing transformer-based models under resource-constrained conditions.

DOI: 10.34148/teknika.v15i1.1431

II. RESEARCH METHODOLOGY

A. Research Flow

This research uses a quantitative experimental approach to design and evaluate a BERT-BiLSTM hybrid model for sentiment analysis of movie reviews using the Rotten Tomatoes Critic Reviews dataset. The initial stage involved collecting movie review text labeled with positive (Fresh) and negative (Rotten) sentiment. This dataset was selected because of its high linguistic diversity, including the use of idioms, informal language style, and varying sentence lengths, which pose challenges for text classification models.

The next stage involves data preprocessing to improve quality before training. This process includes removing non-alphabetic characters and irrelevant symbols, converting all letters to lowercase, and text normalization to reduce unnecessary variation. Next, the text is processed using a BERT tokenizer to convert the sentences into sequences of numeric tokens that the model can understand. Because this model requires uniform input length, padding and truncation are applied according to a predetermined maximum sequence length limit.

Given the very large size of the original dataset and the limitations of CPU-based devices without GPU support, this study used simple random sampling to randomly select a subset of 10,000 reviews using the df.sample(10000, random_state=42) function from the pandas library. This technique ensures each review has an equal chance of being selected, ensuring that the class distribution remains representative of the original dataset. The sampled data was then split into 80% training data and 20% testing data to allow evaluation on data not involved in the training process.

In the development phase, a pre-trained BERT language model was used as a contextual feature extractor, combined with a BiLSTM network as a bidirectional sequence modeler. Preprocessed tokens were fed into BERT to generate contextual embeddings that represent word meanings according to sentence context. These embeddings were then processed by a BiLSTM layer to capture long-range dependencies and word-sequence relationships across both forward and backward directions. The BiLSTM output was then passed through a fully connected layer with a softmax activation function to generate sentiment class probabilities.

The training process was conducted using the AdamW optimizer with a learning rate of 2×10^{-5} for three epochs, and the Cross-Entropy Loss (CEN) loss function measures the difference between model predictions and actual labels. In addition to the hybrid model, this study also trained one BERT model and another BiLSTM model as a benchmark to comprehensively evaluate the effectiveness of the hybrid approach. A low-medium learning rate was used to maintain training stability, while the number of epochs was limited to prevent overfitting and accommodate resource constraints. With this configuration, the total model training time ranged from 4–6 hours for the hybrid model, 2–3 hours for the BERT model alone, and 1 hour for the BiLSTM model alone.

After the training process was complete, all three models were evaluated on test data using metrics such as accuracy, precision, recall, and F1 score to obtain a comprehensive overview of their classification performance. Accuracy represents the proportion of correct predictions across the entire test data set, precision indicates the accuracy of positive predictions, recall measures the model's ability to recognize all true positives, and the F1 score reflects the balance between precision and recall. Furthermore, a confusion matrix is used to analyze the prediction error patterns and model performance for each class in more detail, allowing for a clearer comparison of the relative superiority between the hybrid model and a single model.

After training is complete, the model is evaluated on the test data using the metrics accuracy, precision, recall, and F1 score to provide a comprehensive summary of classification performance. Accuracy indicates the proportion of correct predictions, precision measures the accuracy of positive predictions, recall measures the model's ability to detect all positives, and the F1 score represents the balance between precision and recall. Furthermore, a confusion matrix is used to analyze the distribution of prediction errors and model performance for each class in more detail.

The best performing model is then saved in .pth format for use in the analysis and discussion phase of the results. The entire research process is systematically designed, starting from data collection, pre-processing, model training, to final evaluation, as shown in the research flowchart in Figure 1.

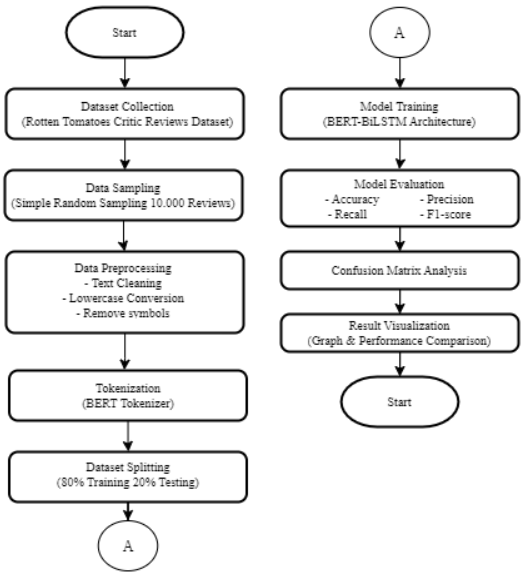

Figure 1. Flowchart of Research Stages

B. Research Dataset

This study uses the Rotten Tomatoes Critic Reviews dataset, a collection of movie reviews compiled from the Rotten Tomatoes website. This dataset contains 1,048,576 reviews from critics across various media outlets. Key attributes in the dataset include rotten_tomatoes_link, critic_name, top_critic, publisher_name, review_type, review_score, review_date, and review_content.

The review_type column is used as a sentiment label, consisting of two categories: Fresh (1) for positive reviews and

Rotten (0) for negative reviews. The class distribution in the original dataset is relatively balanced, making it suitable for binary classification tasks.

However, due to the large dataset size, training a full model using a transformer architecture like BERT requires significant computational resources, particularly in terms of GPU memory and processing time. Therefore, this study uses a subset of 10,000 reviews taken from the original dataset.

The sampling technique used is simple random sampling using the df.sample(10000, random_state = 42) function in the pandas library. This method gives each entry an equal chance of being selected for the sample, thus reducing potential bias. The random_state parameter is used to ensure the sampling process is reproducible in subsequent experiments.

The simple random sampling method was chosen based on the relatively balanced class distribution in the dataset, allowing random sampling to maintain a representative proportion of positive and negative reviews. This approach is also more practical and efficient than other methods, such as stratified sampling, which are typically used when the class distribution is imbalanced.

This sampling strategy was implemented to maintain computational efficiency, given that this study was conducted on a device with an AMD Ryzen 3 5300U CPU without CUDA GPU support. This sample size was deemed sufficient to represent the diversity of data in the original dataset while allowing the training and evaluation processes to run in a reasonable time.

This approach ensures that the data used still reflects the diversity of film critics' opinions and styles, allowing the model to learn complex sentence contexts without losing the key characteristics of the Rotten Tomatoes dataset.

C. Data Pre-Processing

Before the data is used for training, a series of preprocessing steps are performed to ensure the reviewed text meets the BERT model's requirements. The process begins by cleaning the text of non-alphabetic characters, excessive punctuation, and double spaces. All text is then converted to lowercase to ensure consistent word representation.

Once the text is cleaned, the next step is tokenization using BERT tokenization from the bert-base-uncased model. Tokenization involves splitting sentences into token units recognized by BERT, then mapping each token to an index number. The maximum token length is set at 128 tokens, with layers added for short text and truncation for text exceeding the maximum length.

The final preprocessing step is to divide the dataset into two subsets: 80% as the training set and 20% as the test set. The split process is performed randomly using the train_test_split() function to ensure a balanced label distribution between the training and test data.

D. Model Architecture

The model architecture used in this study is a hybrid approach that combines BERT as a contextual embedding generator with BiLSTM as a bidirectional sequential dependency modeler. In this system, BERT acts as the primary feature extractor, transforming each token in the text into a 768-dimensional contextual vector representation. This representation allows the meaning of a word to be determined not only by the word itself but also by its relationship to other words in the sentence. This approach is considered superior to static embedding because it is able to capture context-dependent variations in word meaning (polysemy) and model long-range dependency relationships in text [25], [26].

Conceptually, transformer models like BERT utilize a self-attention mechanism to comprehensively learn relationships between tokens, thus understanding the high-level semantic context within a sentence [25]. Through this mechanism, the model is able to determine the level of relevance of each token to other tokens in a text sequence. Mathematically, this attention mechanism can be expressed as follows in Equation 1:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q is the query, K is the key, and V is the value, while d_k is the dimension of the key vector used to stabilize the normalization process. With this mechanism, BERT is able to generate contextual embeddings that consider the relationships between words in a sentence.

The contextual embeddings generated by BERT are then used as input to the BiLSTM layer, which has 128 hidden memory units. In general, if the input token sequence is expressed as X=(x1,x2,...,xn), then the contextual representation generated by BERT can be written as Equation 2:

$$H = BERT\ (X) \quad (2)$$

where H=(h1,h2,...,hn) represents the contextual embedding that will be processed in the next layer.

BiLSTM processes a sequence of tokens simultaneously in two directions, namely forward and backward, so it can utilize information from both the previous and subsequent contexts in the sentence. This bidirectional approach is crucial in sentiment analysis because the meaning of a word is often influenced by its surrounding context [27]. Theoretically, LSTM is designed to address the vanishing gradient problem in conventional recurrent neural networks through a gating mechanism that regulates the flow of information in long-term memory [27].

The operation of LSTM consists of several main gates that control the flow of information: the forget gate, the input gate, and the output gate. This mechanism can be expressed by the following Equation 3-7:

Forget gate:

$$f_t = \sigma(W_f[h_{t-1,x_t}] + b_f) \quad (3)$$

Input gate:

$$i_t = \sigma(W_i[h_{t-1,x_t}] + b_1) \quad (4)$$

Candidate cell state:

$$\tilde{C}_t = tanh(W_c[h_{t-1,x_t}] + b_c) \quad (5)$$

Cell state update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (6)$$

Hidden state:

$$h_t = o_t \cdot \tanh(C_t) \quad (7)$$

In the BiLSTM model, two LSTM processes are run in parallel in the forward and backward directions, resulting in the following combined representation (Equation 8):

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (8)$$

The combination of BERT and BiLSTM allows the model to leverage the global context understanding of transformers and the sequence modeling capabilities of recurrent neural networks [26], [27].

The output from the BiLSTM layer is then passed to a dropout layer with a value of 0.3 to reduce the risk of overfitting by randomly deactivating some neurons during the training process. The results are then processed by a fully connected layer consisting of two neurons to represent two sentiment classes. The classification stage is performed using a softmax activation function to generate probabilities for positive and negative sentiment classes, which can be mathematically written as (Formula 9):

$$y = softmax(Wh_t + b) \quad (9)$$

where y is the sentiment class probability, W is the model weight, and b is the bias.

While this hybrid architecture (Figure 2) has the potential to improve classification performance, high model complexity can also increase the risk of overfitting, especially when the amount of training data is relatively limited compared to the number of model parameters. Overfitting occurs when a model overfits to the training data, resulting in degraded performance when tested on new data [26]. In the BERT-BiLSTM model, this risk arises because BERT has a very large number of pre-trained parameters, and the addition of BiLSTM layers further increases model complexity during the fine-tuning process.

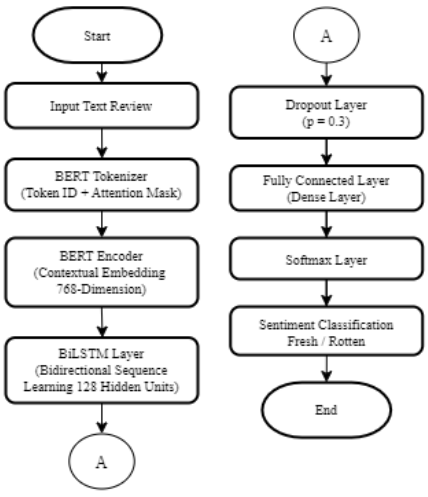

Figure 2. Flowchart of the Architecture

E. Training Configuration

The model was trained using Visual Studio Code (VS Code) as the development platform with Python 3.12, PyTorch 2.3.0, and Transformers 4.41.0. Because the hardware used did not support CUDA GPUs, the entire training process was run on an AMD Ryzen 3 5300U CPU (4 cores, 8 threads, 2.6 GHz) with Radeon Graphics (512 MB dedicated, 3.7 GB shared).

Training was performed using the Adam optimizer and a learning rate of 2e-5. The batch size was set to 16, the number of epochs was three during the full training run on the training data, and the loss function used was CrossEntropyLoss. A rate dropout of 0.3 was applied to maintain model generalization. The most optimal model was then saved in model_bert_bilstm.pth format using the torch.save() function. This training configuration was chosen to balance training stability and computational efficiency, taking into account the limitations of the hardware used. Even though the training is done without CUDA GPU, the model results are still academically valid because the entire forward and backward propagation process still runs entirely on the CPU.

F. Model Evaluation

The evaluation phase involves testing the model on a test dataset, which comprises a total of 20% of the data. The assessment includes four key metrics: accuracy, precision, recall, and F1 score. Accuracy measures the percentage of predictions that are close to correct based on the test data, while precision describes the proportion of the system's true positive predictions out of all true positive predictions. Recall indicates the model's ability to recognize all true positive data, and the F1 score is used as a harmonic mean between recall and precision, balancing the two.

The confusion matrix results show that the model tends to predict the Fresh class more often than the Rotten class. This can be influenced by several factors. One is the data distribution in the subset used during the training process. Although the original dataset has a relatively balanced class distribution, the random sampling process can potentially produce slightly different class proportions, allowing the model to learn more patterns from certain classes. Furthermore, reviews with positive sentiment are easier for the model to recognize. Conversely, reviews with negative sentiment are often conveyed indirectly or contain a combination of positive and negative expressions in a single sentence, which makes the classification process more challenging. Another factor that may influence this is the self-attention mechanism in transformer-based models like BERT, which tends to give greater weight to words with strong sentiment associations, thus increasing the likelihood of predicting the more dominant class in the training data.

In addition to numerical parameter calculations, the evaluation also uses a Confusion Matrix to explain the distribution of FALSE and TRUE predictions between the two classes (New and Rotten). To clarify the model's performance, a bar graph of the parameter evaluation was created using the Matplotlib library, displaying the percentages of accuracy, precision, recall, and F1-score simultaneously.

G. Implementation Flow

The overall research process begins with loading and cleaning the dataset, followed by text tokenization using BERT tokenization. The data is then divided into a training dataset and a testing dataset. A BERT-BiLSTM model is created and trained using Adam optimization with a learning rate of 2e-5. After training, the model is tested on the required test data to gauge its performance. Finally, the best model is saved as a .pth file for future analysis and replication.

With this methodology, the BERT-BiLSTM model is expected to be able to simultaneously learn semantic context and sequential stable relationships, resulting in stable and accurate performance in analyzing movie review sentiment on the Rotten Tomatoes dataset.

III. RESULT

A. Model Training Results

BERT-BiLSTM was trained using 10,000 movie review data samples randomly sampled from the Rotten Tomatoes Critic Reviews dataset. The model was trained for two epochs with a large subset size of 16, a learning rate of 2e-5, and the Adam Optimizer. Training was performed entirely on an AMD Ryzen 3 5300U CPU without CUDA GPU acceleration.

During the training process, the loss function value decreased, indicating that the model had successfully learned sentiment patterns from the data. The loss value in the first epoch was recorded at 42%, decreased to 33% in the second epoch, and decreased again in the third epoch to 27%, indicating training convergence.

The trained model was then saved in model_bert_bilstm.pth format for use in the evaluation phase. In addition to the hybrid model, a single BERT model and a single BiLSTM model were also trained separately for comparison to assess the effectiveness of the proposed hybrid approach. Total training time ranges from 4-6 hours, which is relatively long due to hardware limitations, such as computing using only the CPU without GPU acceleration. The architectural implementation involves a BERT transformer model with a very large number of parameters, making each computational process on the dataset more demanding than conventional deep learning models.

Training is conducted using a data subset of 10,000 reviews with a maximum sequence length of 128 tokens per sample. Although the dataset size is reduced through random sampling to improve efficiency, BERT tokenization, embedding formation, and sequence processing still require significant time when running on the CPU. To accommodate these limitations, training parameters are adjusted, including using a limited number of epochs and a small batch size. These adjustments aim to reduce the computational load and speed up the training process without significantly reducing model performance. The same approach is applied to training the benchmark model to ensure fair evaluation performance among models under equivalent computational conditions.

With this configuration, the training time can be maintained within the range of 4-6 hours, which is still considered reasonable for training a transformer-based model without GPU support. If training is done using GPU, the time required can be significantly reduced to only a few tens of minutes.

B. Model Evaluation Results

The evaluation was conducted using 20% of the test data separated from the initial dataset to assess the model's ability to classify movie reviews into two sentiment categories: Fresh (positive) and Rotten (negative). The BERT-BiLSTM model testing results are shown in Table 1 as a confusion matrix. Based on the table, the model successfully classified 1,111 Fresh reviews and 152 Rotten reviews correctly. However, there were still prediction errors, with 566 Rotten reviews classified as Fresh and 171 Fresh reviews classified as Rotten. This finding indicates that the model has a tendency to predict the positive class more often than the negative class. The distribution of these predictions is visualized in Figure 3, which shows the dominance of prediction results in the Fresh class.

Table 1 BERT-BiLSTM Model Confusion Matrix

	Fresh Prediction	Rotten Prediction
Actual Fresh	1,111	152
Actual Rotten	171	566

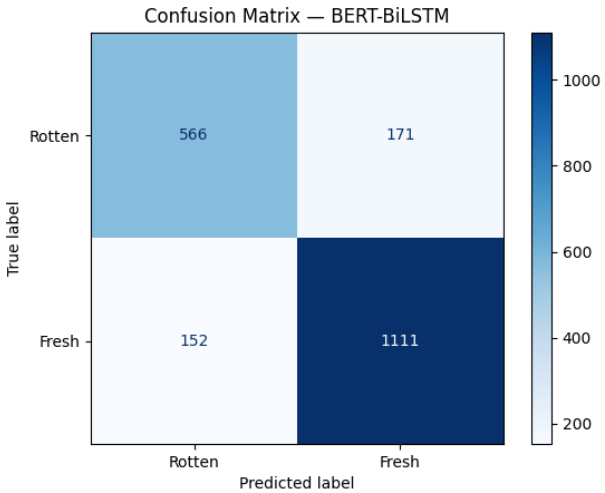

Figure 3. BERT - BiLSTM Model Confusion Matrix

These parameters, based on the retraining results of the hybrid model, showed significant improvement. The BERT-BiLSTM model evaluation can be seen in Table 2. Based on these results, the model achieved significant similarity to the evaluation results. The results showed relatively similar accuracy, precision, recall, and F1-score values, namely 84%. The accuracy value reached 84% indicating that the majority of the review data in the testing phase was successfully classified correctly. The precision reached 84% indicating that the sentiment predictions generated by this model are of good quality. The accuracy level, while the recall reached 84%,

indicates the model's ability to detect most of the relevant data. Finally, the F1-score value of 84% indicates an optimal balance between precision and recall, so the model is considered not only accurate but consistent in recognizing sentiment patterns. This uniformity certainly indicates that this hybrid model demonstrates stable performance without any bias towards a particular class. It is also capable of capturing global context and word order effectively in fairly complex review texts. However, the text model's performance still has potential for improvement through extensive refinement, the use of a larger dataset, and adequate computational support to optimize model generalization.

Experimental results show that the BERT-BiLSTM model is capable of delivering better performance than the single model used as a comparison. Combining BERT's contextual representation with BiLSTM's sequential modeling capabilities allows the model to capture more complex sentiment patterns in movie reviews. The Rotten Tomatoes dataset has quite diverse text characteristics, such as the use of metaphors, irony, and long sentences that often contain mixed sentiments. In these situations, using a single model like BERT or BiLSTM alone is often insufficient to understand the full context and relationships between words in a sentence.

By combining these two approaches, the BERT-BiLSTM model can produce richer and more comprehensive feature representations, thereby improving the accuracy of sentiment classification. This improvement is reflected in higher evaluation scores compared to single models on the same dataset.

Table 2. Evaluation Parameters of the BERT-BiLSTM Model

Parameter	Value
Accuracy	84%
Precision	84%
Recall	84%
F1-score	84%

The evaluation results of the BERT model without BiLSTM integration are presented in Table 3. This model successfully classified 84 "Bad" reviews and 1,233 "Fresh" reviews correctly, with a lower misclassification rate than the hybrid model. The confusion matrix visualization in Figure 4 shows that the distribution of predictions between the two sentiment classes is more balanced.

Table 3. BERT Model Confusion Matrix

	Fresh Prediction	Rotten Prediction
Actual Fresh	1,233	30
Actual Rotten	653	84

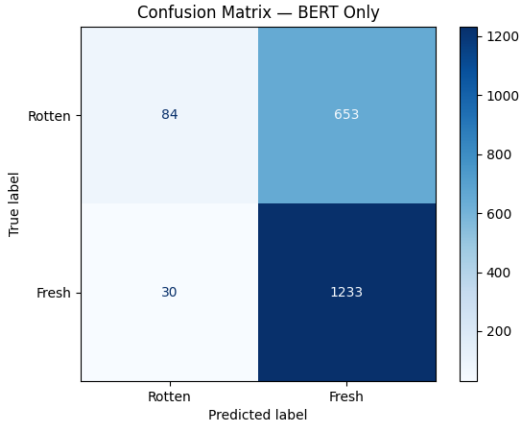

Figure 4. BERT Model Confusion Matrix

The BERT model parameter evaluation is shown in Table 4, where the BERT-only model showed an accuracy of 66%, meaning that the model's accuracy is still in the medium category, tending to be lower than the BERT-LSTM hybrid model with a result of 84%. The accuracy value of 68% indicates that it produces fairly accurate sentiment predictions, but is still not as accurate as the hybrid model in minimizing classification errors, considering that 66% indicates that the model has not been able to optimally detect all sentiment data, resulting in many classification errors from the review. Finally, the F1 score of 57% indicates its balance in accuracy and relatively low recall. This indicates that the overall model is still unstable. This certainly strengthens the role of BiLSTM in the hybrid model, which makes a significant difference in capturing the dependence of word order and local context, resulting in more accurate and balanced sentiment classification.

Table 4. BERT Model Evaluation Parameters

Parameter	Value
Accuracy	66%
Precision	68%
Recall	66%
F1-score	57%

Meanwhile, the evaluation results of the BiLSTM model without BERT integration, as shown in Table 5, indicate that the model is unable to accurately classify the Rotten class, as all Rotten samples are predicted as fresh. The visualization in Figure 5 also shows the dominance of predictions from only one class, indicating a tendency for the model to be biased towards the positive class.

Table 5. Confusion Matrix of BiLSTM Model

	Fresh Prediction	Rotten Prediction
Actual Fresh	1,263	0
Actual Rotten	737	0

DOI: 10.34148/teknika.v15i1.1431

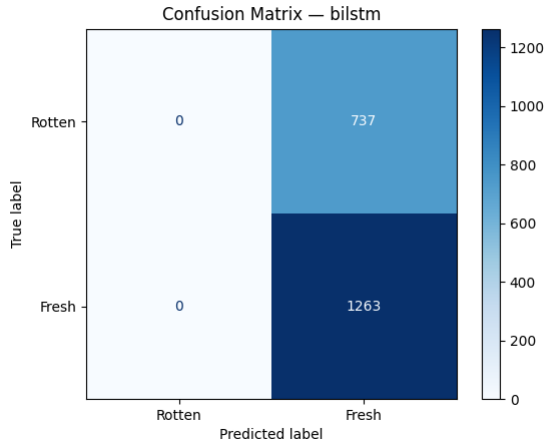

Figure 5. BiLSTM Model Confusion Matrix

The parameter evaluation of the single BiLSTM model presented in Table 6 shows an accuracy of 63%, lower than the two models described previously, but still in the moderate category in its classification ability. The accuracy of 39% is quite low, indicating that many of the model's predictions are partially inaccurate, thus leading to a high classification error rate. The recall of 63% indicates that the model is capable of finding relevant sentiment data, but is still not as accurate as the previous model. Meanwhile, the F1 score of 48% is quite low, reflecting the balance between precision and recall, so the overall performance is considered unstable. The low evaluation results of the single BiLSTM model indicate that without the support of contextual representation according to BERT, the model has difficulty capturing implicit meaning, understanding long sentence structures, and understanding global context in movie reviews. Therefore, a hybrid model is fundamental to address this issue. Combining these two models will produce good performance in systems that have been created in the context of movie review prediction.

Table 6. BiLSTM Model Evaluation Parameters

Parameters	Value
Accuracy	63%
Precision	39%
Recall	63%
F1-score	48%

A comparison of the overall performances of the three models is presented in Table 7. The BERT-BiLSTM model showed the best performance, with accuracy, precision, recall, and F1-score values reaching 84% each, surpassing both the single BERT and single BiLSTM models. The BERT model ranked second with an accuracy of 66% and an F1-score of 57%, while the BiLSTM model performed the lowest among the three. These findings indicate that the hybrid approach is capable of integrating the contextual representation strengths of BERT with the sequence dependency modeling capabilities of BiLSTM, providing a better understanding of sentiment than either model alone.

Table 7 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
BERT-BiLSTM	84%	84%	84%	84%
BERT	66%	68%	66%	57%
BiLSTM	63%	39%	63%	48%

IV. DISCUSSION

Retraining results show significant performance differences between BERT-BiLSTM, single BERT, and single BiLSTM models. The comparison presented in Table 7 shows that the BERT-BiLSTM model achieved the highest performance, with accuracy, precision, recall, and F1-score values of 84% each. The single BERT model ranked second with 66% accuracy and a 57% F1-score, while the single BiLSTM model performed the lowest with 63% accuracy and a 48% F1-score. These results indicate that combining BERT's contextual representation with BiLSTM's bidirectional sequence modeling can improve sentiment classification effectiveness compared to using each model separately.

The superiority of the BERT-BiLSTM model demonstrates that the integration of the transformer and recurrent neural network architectures effectively captures both global context and local sequential relationships among words. Movie critics' reviews often contain implicit meanings, metaphors, and long sentence structures that require an understanding of broad context and sequence dependencies. BERT, through its self-attention mechanism, produces rich semantic representations, while BiLSTM models the dynamics of word sequences in both forward and backward directions, enhancing the model's ability to understand the development of sentiment within a sentence.

A single BERT model still performs well thanks to its ability to capture the global context of relationships between tokens. However, without an additional sequence modeling layer, this model is limited in representing complex local dependencies, especially in long sentences with nested syntactic structures. This explains why its performance remains lower than the hybrid model, although still superior to the single BiLSTM model.

In contrast, the single BiLSTM model performed poorly because it relied solely on static embeddings without supporting global context understanding. Without contextual representations like those provided by BERT, the model struggled to capture complex language nuances, especially in reviews with implicit sentiment or long sentence structures. Low precision and F1-scores indicated a higher frequency of misclassification errors in this model. Evaluation results showed that the single BiLSTM model classified all data into the Fresh class, indicating that it failed to effectively distinguish between the two sentiment classes. This condition can be interpreted as an extreme form of prediction bias or model collapse, where the model fails to learn distinguishing patterns between classes and ultimately tends to favor one

DOI: 10.34148/teknika.v15i1.1431

dominant class to minimize the loss function value during the training process.

One contributing factor to this condition is BiLSTM's limited ability to understand complex semantic context. This is because the model only utilizes basic embedding representations without the support of more robust contextual representations like those found in transformer-based models. Furthermore, the sentence structure in movie reviews is generally long, ambiguous, and often contains mixed sentiments, requiring a deeper understanding of context. Without adequate context representation capabilities, the BiLSTM model is less able to capture the relationships between words that determine sentiment polarity. As a result, the optimization process tends to produce imbalanced predictions and lower accuracy.

The method proposed in this study has several advantages in classifying sentiment in movie reviews. The combination of BERT and BiLSTM allows the model to leverage the contextual representation power of the transformer architecture and the sequence modeling capabilities of recurrent neural networks. BERT plays a role in understanding word meaning based on the overall context of a sentence, while BiLSTM is able to capture sequential relationships between tokens in both directions. The integration of these two approaches allows the model to produce a more comprehensive feature representation, thus improving classification accuracy compared to using a single model.

Although the BERT-BiLSTM model demonstrated superior performance compared to the benchmark model, this approach still has several limitations. One of these is the relatively high architectural complexity due to combining the transformer model with a recurrent neural network, which requires greater computational resources. Furthermore, the training process, conducted without GPU support, limits hyperparameter exploration and the number of epochs that can be applied, thus limiting the model's performance potential. Another limitation is the use of a subset of the dataset, consisting of 10,000 samples, which likely does not fully reflect the diversity of data contained in the Rotten Tomatoes dataset as a whole.

Overall, this study confirms that transformer-based contextual representations play a key role in modern sentiment analysis, while integration with sequential architectures like BiLSTM can provide significant performance improvements. In practical applications, such models have potential applications for analyzing public opinion on movies, products, or services, monitoring social media sentiment, review-based recommendation systems, and data-driven decision-making in the entertainment and digital marketing industries.

V. CONCLUSION

This study successfully designed, implemented, and evaluated a BERT-BiLSTM hybrid model for sentiment classification on the Rotten Tomatoes Critic Reviews dataset, while comparing it with both a single BERT and a single BiLSTM model. The integration of BERT as a deep contextual representation generator with BiLSTM as a bidirectional sequence dependence modeler allows the hybrid model to simultaneously capture global context and local sequence patterns in complex movie review texts. Conceptually, this approach provides a comprehensive framework for understanding the nuances of language, implicit opinions, and long sentence structures commonly found in critical reviews.

Experimental results show that the BERT-BiLSTM model performed the highest with accuracy, precision, recall, and F1-score of 84% each, followed by the single BERT model with 66% accuracy and 57% F1-score, and the single BiLSTM model with 63% accuracy and 48% F1-score. These findings suggest that the integration of transformer-based contextual representation with bidirectional sequence modeling can improve the effectiveness of sentiment analysis compared to using each model separately. This combination is proven to simultaneously benefit from BERT's global context understanding and BiLSTM's sensitivity to word order.

From a practical perspective, the BERT-BiLSTM hybrid model has potential applications in public opinion analysis, social media monitoring, service evaluation, recommendation systems, and product and movie review analysis, all of which require deep contextual understanding. For future research, the use of larger datasets, support for GPU acceleration, and the development of more efficient hybrid architectures are recommended to optimize the benefits of the combination of transformers and recurrent neural networks. Overall, this research demonstrates that the BERT-BiLSTM hybrid approach is an effective and promising solution for sentiment analysis of complex text.

REFERENCES

[1] K. Chakraborty, S. Bhattacharyya, and R. Bag, “A Survey of Sentiment Analysis: Approaches, Datasets, and Trends,” *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 450-464, 2020, [Online]. Available: https://ieeexplore.ieee.org/

[2] F. Aftab and et al., “A Comprehensive Survey on Sentiment Analysis Techniques,” *International Journal of Technology (IJTech)*, 2022, [Online]. Available: https://ijtech.eng.ui.ac.id

[3] S. Jain, “Sentiment Analysis,” in *Proceedings of the ACL Workshop*, 2017. [Online]. Available: https://aclanthology.org

[4] “Review of Static vs Contextual Word Embeddings in NLP,” *HighTech Journal*, 2020, [Online]. Available: https://hightechjournal.org

[5] “BERT in Sentiment Analysis: A Review,” *HighTech Journal*, 2021, [Online]. Available: https://hightechjournal.org

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL-HLT*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[7] J. Joseph, "A Survey on Deep Learning Based Sentiment Analysis," *ScienceDirect*, 2022, [Online]. Available: https://www.sciencedirect.com
[8] S. Alaparthi and M. Mishra, "BERT: A Review of Applications in Sentiment Analysis," *arXiv Preprint*, 2020, [Online]. Available: https://www.researchgate.net/
[9] "Polarity of Yelp Reviews: A BERT-LSTM Comparative Study," *MDPI*, 2024, [Online]. Available: https://www.mdpi.com
[10] "BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Reviews," *MDPI*, 2022, [Online]. Available: https://www.mdpi.com
[11] A. S. Talaat and et al., "Sentiment Analysis Classification System Using Hybrid BERT Models," *J. Big Data*, 2023, [Online]. Available: https://journalofbigdata.springeropen.com
[12] "LSTM and BiLSTM in Sequential Text Processing: Concepts and Applications," in *Proceedings of ACL Workshop on Deep Learning for NLP*, 2019. [Online]. Available: https://aclanthology.org
[13] N. Sikana, S. Winardi, G. -, G. F. Situmorang, and R. Lubis, "Analisis Sentimen untuk Ulasan Produk E-Commerce Shopee Menggunakan BERT," *Jurnal Sifo Mikroskil*, vol. 26, no. 2, Oct. 2025, doi: 10.55601/jsm.v26i2.1796.
[14] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The hybrid of BERT and deep learning models for Indonesian sentiment analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 591-602, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.
[15] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 3, p. 1018, Jul. 2021, doi: 10.30865/mib.v5i3.3111.
[16] "Sentiment Analysis Using FastText and LSTM: A Comparative Study," *ResearchGate Preprint*, 2020, [Online]. Available: https://www.researchgate.net
[17] "BERT for Sentiment Analysis on Rotten Tomatoes Reviews," *ResearchGate*, 2021, [Online]. Available: https://www.researchgate.net
[18] "Hybrid BERT + LSTM and BERT + CNN Models for Sentiment Analysis on Social Platforms," *UAD Repository*, 2022, [Online]. Available: https://eprints.uad.ac.id
[19] "Fine-Tuning and Compression Strategies for Practical BERT Deployment," *arXiv Preprint*, 2022, [Online]. Available: https://arxiv.org
[20] "Lightweight BERT Models for CPU-Only Environments," *SpringerOpen Journal*, 2023, [Online]. Available: https://www.springeropen.com
[21] "Comprehensive Review on BERT-based Sentiment Analysis Techniques," *arXiv Preprint*, 2021, [Online]. Available: https://arxiv.org
[22] "Film Review Datasets and Their Characteristics for Sentiment Analysis," *Science Gate Journal*, 2020, [Online]. Available: https://science-gate.com
[23] "Static vs Contextual Embeddings: A Comparative Study for Word Representation," *HighTech Journal*, 2021, [Online]. Available: https://hightechjournal.org
[24] "Comparative Deep Learning Models for Film Review Sentiment Analysis," *Science Gate Journal*, 2021, [Online]. Available: https://science-gate.com
[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Online. [Online]. Available: https://huggingface.co/
[26] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, pp. 103694-103704, 2022, doi: 10.1109/ACCESS.2022.3210182.
[27] A. Chowanda and Y. Muliono, "Indonesian Sentiment Analysis Model from Social Media by Stacking BERT and BI-LSTM," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IEEE, Sep. 2022, pp. 278-282. doi: 10.1109/AiDAS56890.2022.9918717.