Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

A Comparative Study of BERT, RoBERTa, and Bi-LSTM for Sentiment Analysis of Social Media Data in Diverse Domains

Sunny Suresh Nahar, Swapna Kulkarni, Dr. Prashant Prakashrao Agnihotri
School of Technology, Swami Ramanand Teerth Marathwada University, Sub-Campus, Latur, Maharashtra, India

Abstract **– Sentiment Analysis (SA) has become a very critical activity for gaining valuable information from the huge amount of user-generated content that is available on social media sites. Yet, sentiment classification has proved to be difficult due to reasons like language informality, contextual uncertainty, noise in the text, abbreviations, emojis, etc. The goal of this research is to conduct a comparative assessment of Bi-directional Encoder Representations Transformer (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) models for SA on various social media domains through one common experimental setup. The suggested experimental setup would include text processing, feature extraction using Word2Vec, feature selection using Mutual Information (MI), and training and testing of the models on the Sentiment140 Dataset with 1.6 million tweets. Results from experiments indicate that the RoBERTa-based framework attains the highest accuracy of 98.56% along with the highest precision, recall, F1 score, and 5.46 seconds of training time and outperforms Bi-LSTM and BERT in terms of classification efficacy and computational efficiency. The novelty of the proposed study lies in ensuring a fair comparison through identical experimental settings while integrating Word2Vec-based feature representation with Mutual Information-based feature selection to enhance classification performance. The suggested framework is a highly reliable and scalable solution for performing accurate sentiment classification and helps in selecting the right model.**

Keywords – BERT; Bi-LSTM; Mutual Information; RoBERTa; Sentiment Analysis

I. INTRODUCTION

The rapid surge of social media utilization has yielded an unprecedented volume of user-generated content, conveying public opinions within multiple sectors, including medicine, politics, finance, marketing, and entertainment [1], [2]. Consequently, SA has become an important task that can help to extract useful information from such content for an organization to track public opinion or to forecast trends [3-5]. Nonetheless, social media text SA has great difficulties. The use of informal and noisy content such as slang, abbreviations, emojis, and poor grammar is commonly seen in social media content [6], [7]. The performance of the model gets affected due to the high variability in domains. Sarcasm, contextual ambiguity, multilingual or code-mixed inputs, class imbalance, etc., can severely affect classification robustness [8], [9]. Models that can capture deep, appropriate semantics and long -range dependencies are needed to tackle this complexity [10], [11]. Fig. 1 illustrates the basic steps of SA.

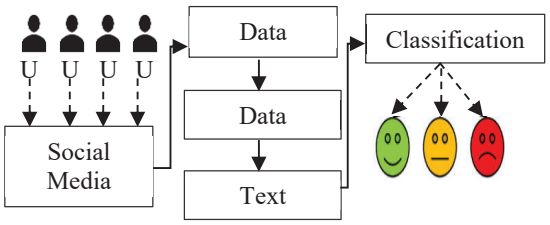


Fig. 1. Basic steps of SA [12]

Traditional Machine Learning (ML) algorithms [13], [14] that adopt Bag-of-Words and TF-IDF rely heavily on hand-crafted features, despite being more efficient than other types of models [15-18]. Nevertheless, they cannot integrate contextual relations, word dependency owing to order, and circa semantics; hence, they do not generalize well across domains of different types [19-21]. Deep Learning's (DL) recent advancements have substantially enhanced the performance of sentiment categorization tasks, as an expansion of Bi-LSTM networks [22], which may learn the dependencies in both directions at the same time [23], [24]. Transformers' architectural design has introduced new possibilities for natural language processing through its dual mechanisms of self-attention and pretraining, which enable better contextual comprehension and domain adaptation abilities for users [25], [26]. The different architectural design elements, which include their capacity to represent context, their total number of parameters, and their processing needs, can determine how well they perform with various data sets [27]. The existing research requires domain experts to understand different model selection methods because professional fields require accuracy and generalization, and computational costs to be balanced [28]. The current study carries out an extensive evaluation that compares the performance of the three DL models for social media sentiment data analysis in different domains. The study gives a comprehensive assessment of performance metrics, generalization ability, and computational speed on various social media platforms, which is different from other effort that employs one architecture on domain-specific datasets. This research proposes an evaluation methodology that directly compares Bi-LSTM models under a standardized test system, along with the transformer-based models, BERT and RoBERTa, on multiple dimensions. Comparison between recurrent networks and self-attention networks are possible from identical setups and evaluations, which reveal how each one learns context. This research aims to compare the BERT and RoBERTa models that analyze the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

sentiment of social media data across different industries. This research includes these essential contributions:

An extensive empirical study of Bi-LSTM, BERT, and RoBERTa models for multi-domain sentiment classification.

Establishing a common testing platform to measure the robustness and generalization capabilities of the models on various social media platforms.

Testing the trade-offs between the models' performance and efficiency, including accuracy, training time, and complexity.

Offering valuable recommendations for model selection in real-world SA tasks.

This paper is organized as follows. Section 2 gives an overview of the related literature. The data and techniques used, and the steps followed, are covered in Section 3. Section 4 presented the results and discussion of the research. Lastly, Section 5 shows the conclusion and future directions for research.

II. LITERATURE REVIEW

In this section, some research works on SA using social media data in different fields are discussed. In 2026, Arularasan et al. [29] presented a new Cycle Generative Adversarial Network (GAN) based Term Memory integrated with SA (CyG-TMSA) approach to improve the performance of opinion mining tasks on social media datasets. The results revealed that the proposed approach achieved 98% accuracy with a processing time of 10.12 s on Amazon and Twitter datasets. Sofi et al. (2026) [30] presented an adaptive ensemble ML classifier for Twitter-based SA by combining four classifiers using voting, bagging, Extreme Gradient Boosting (XGBoost), and stacking with a meta-learner, which performed better than the standalone classifiers in terms of accuracy and robustness. Rahman et al. (2025) [31] combined the hybrid RoBERTa-BiLSTM model to capture remote dependencies. The results of the experiment proved to be successful, as the accuracy for Twitter US Airline was 80.74%, 92.36% for IMDB, and 82.25% for Sentiment140. Siddiqui et al. (2025) [32] conducted a study that comparatively assessed ML and hybrid Bi-LSTM with Word2Vec approach on the Sentiment140 dataset, and it found that the hybrid Bi-LSTM architecture outclassed traditional ML models in terms of overall classification accuracy based on confusion matrix evaluation. Deng and Liu (2025) [33] proposed a RoBERTa–BiLSTM–Attention Model incorporated with a supplementary sentiment dictionary based on SO-PMI. Their model achieved 91.56% accuracy, 90.87% precision, 91.07% recall, and 91.17% F1-score, suggesting it classification and generalization capability was improved. Karim et al. in 2025 [34] applied support vector machines, Naïve Bayes, and LSTMs in addition to BERT on the Sentiment140 dataset for SA. The results showed that BERT achieved the best accuracy of 83.48 for predicting trends in consumer behavior.

Employing the IMDb and Sentiment140 datasets, Elmitwalli and Mehegan (2024) [35] conducted a comprehensive evaluation of several SA models, including lexicon-based, ML, DL, and artificial intelligence. They found that the highest F1-scores were achieved by BERT (0.9380 on IMDB; 0.8114 on Sentiment140), while GPT-3 provided the best overall generalization across those partially unannotated COP9 tweets (0.8812). The Dual Joint Classifier (DJC) -based model that Ram (2024) [36] suggested for balancing classes in Twitter SA is a combination of RoBERTa and BERT architectures, with layers of Bidirectional-Gated Recurrent Unit (Bi-GRU) and Bi-LSTM. The DJC model achieved 87.22% accuracy with the US Airline Twitter data set and 93.87% accuracy with the Apple Twitter data set. Tang (2023) [37] created a Transformer-based deep neural network model to perform sentiment classification on the 1.6 million tweets provided by the Sentiment140 dataset. Experimental results demonstrated significant accuracy and robustness improvements, indicating that Transformer architectures have a better capability to handle contextual dependencies with large-scale social media data. The Bi-LSTM-based methodology given by Arbane et al. (2023) [38] was utilized to classify public sentiment regarding COVID-19 technology. Additionally, Tan et al. (2022) [39] proposed a hybrid RoBERTa-LSTM architecture that integrated the transformational model's contextually embedded information with sequentially created LSTM architectures and achieved F1 results of 93% for IMDb, 91% for Twitter US Airline, and 90% for Sentiment140.

However, some research gaps have been identified despite breakthroughs in the field of SA. Most hybrid and ensemble models were evaluated on standard datasets but were not thoroughly analyzed for cross-domain capabilities and/or evaluated for real-world use [29], [34], [39]. Although BERT and GPT-3 models were noted for better handling of tasks, the complexity and the need for Real-time streaming dealt with the models' poor capability [35], [36]. Also, some of the main problems have not been analyzed in-depth, like sarcastic speech detection, multilingual sentiment, low-resource language sentiment, and refined emotional sentiment [32], [38]. Lastly, explainable and interpretable AI was not analyzed at all. This is one of the most important factors for Trust in AI (TnAI) and is significant for improving Trust and Transparency in the sentiment decision support systems.

III. RESEARCH METHODOLOGY

This section shows the proposed method for multimodal SA, which uses text data for precise classification results, as shown in Fig. 2. The collected data is preprocessed using traditional text processing techniques such as tokenization, removal of stop-words, lemmatization, and normalization. As a feature extraction method, Word2Vec is used with MI for feature selection, allowing identification of the most important features. Bi-LSTM, BERT, and RoBERTa are trained earlier after splitting the data into training and testing data. The training process uses the cross-entropy loss function and Adam's optimizer. proposed model performance is evaluated using conventional evaluation metrics and is monitored for its steadiness and elasticity for the entire duration of the monitoring process.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

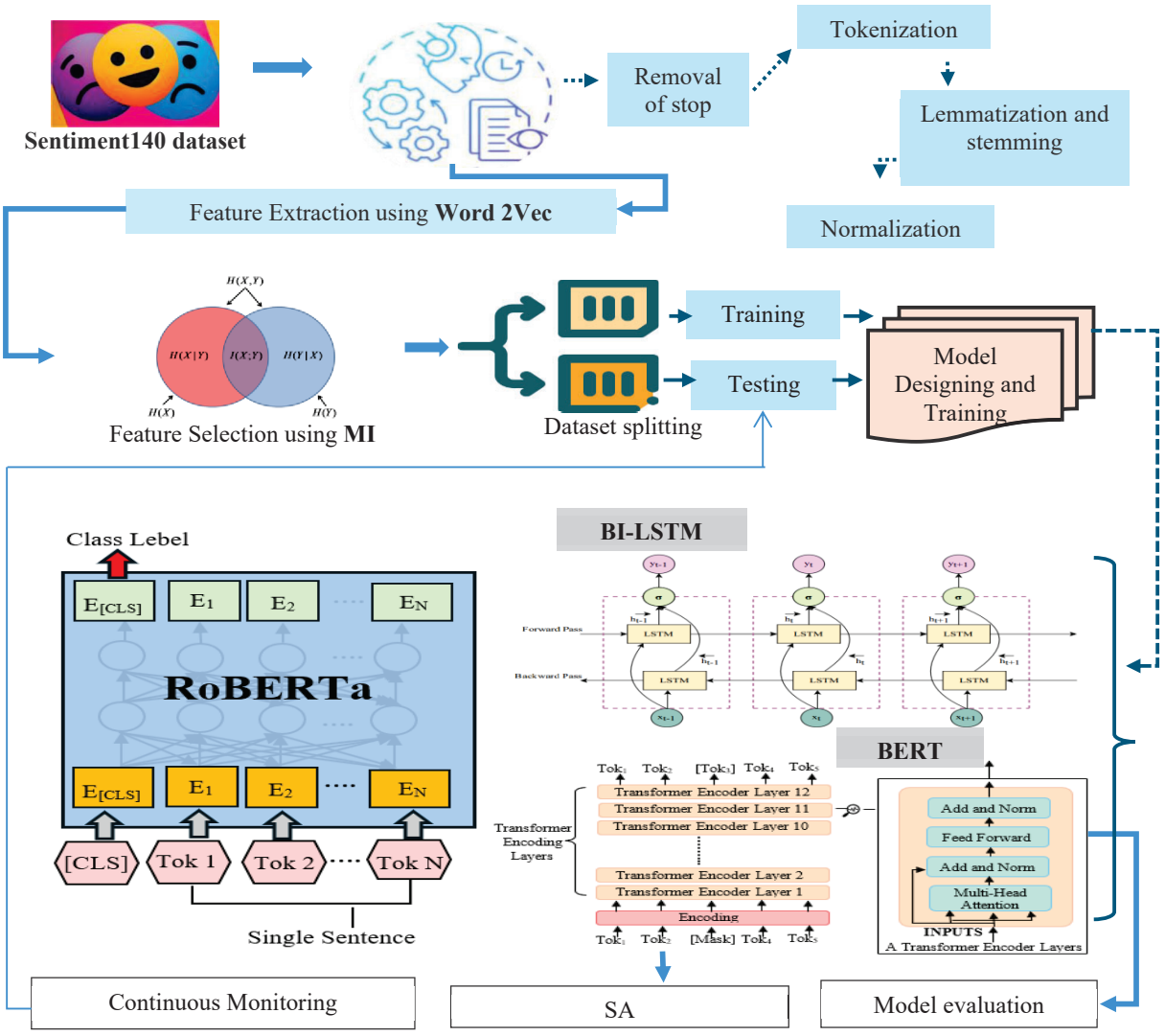


Fig. 2. Design Architecture

A. *Dataset Explanation*

The dataset for SA in this study comes from the open-source Kaggle [40] known as Sentiment140, which is specifically meant for the task of SA on Twitter data at scale. The dataset contains roughly 1.6 million tweets, all of which have been automatically labelled in terms of polarity in sentiments and thus are suitable for binary classification tasks. Each data entry contains features such as tweet ID, date, username, query feature, and tweet text. This dataset captures real-life usage of social media in terms of language, emoticons, abbreviations, and other noisy features, thus making it a great testbed for evaluating the performance of DL and transformer-based SA methods. Table I displays the structure of the dataset used for training.

TABLE I: DATASET SUMMARY

Feature Name	Description	Example
Total Records	Number of tweets	1,600,000
Target	Sentiment polarity (0 = Negative, 4 = Positive)	0 / 4
Tweet_ID	ID for tweet	1467858369
User	Username	username123
Tweet_Text	Raw tweet content	“I love this movie :).”

B. *Data Preprocessing*

To convert noisy and unstructured social media language into a clear, consistent format appropriate for DL models, data preprocessing is an essential step in SA [41], [42]. To convert the unprocessed multimodal textual information into a clear, organized representation that can be used for feature removal, the preprocessing step is carried out in the suggested technique. Let the raw textual corpus be denoted as $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$, where each document $d_i$consists of a sequence of words d_i. The first step is the removal of stop words, which is performed to eliminate commonly occurring but semantically insignificant words (e.g., “is”, “the”, and “and”). Let $\mathcal{S}$ denote the predefined stop-word set. The filtered token set $\mathcal{T}'(d_i)$ is obtained as:

$$\mathcal{T}'(d_i) = \{t_k \in \mathcal{T}(d_i) \mid t_k \notin \mathcal{S}\} \quad (1)$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

This operation reduced dimensionality and improved discriminative feature learning.
Next, tokenization, in which each document is segmented into individual lexical units (tokens). Formally, a tokenization function $\mathcal{T}(\cdot)$ is applied as:

$$\mathcal{T}(d_i) = \{t_1, t_2, \dots, t_K\}, t_k \in \mathcal{V} \quad (2)$$

Where $\mathcal{V}$ is the vocabulary space and $K \leq N$ after cleaning punctuation and delimiters.
Subsequently, lemmatization and stemming were applied to reduce morphological variations. For stemming, a function $\mathcal{F}_s(\cdot)$truncated tokens to their root forms:

$$t_k^{(s)} = \mathcal{F}_s(t_k) \quad (3)$$

Where $t_k^{(s)}$represents the stemmed version of token t_k. In contrast, lemmatization used a vocabulary-based mapping function $\mathcal{F}_l(\cdot)$to obtain the canonical lemma:

$$t_k^{(l)} = \mathcal{F}_l(t_k \mid \text{POS}) \quad (4)$$

Where POS denotes the part-of-speech tag, ensuring linguistically valid base forms (e.g., "running" → "run"). Finally, normalization was conducted to standardize token representations. This included lowercasing, removal of special characters, and numerical standardization. Formally, a normalization operator $\mathcal{N}(\cdot)$was applied:

$$t_k^{(n)} = \mathcal{N}(t_k) \quad (5)$$

Where typically:

$$\mathcal{N}(t_k) = \text{lowercase}(t_k), t_k \in \mathcal{V} \quad (6)$$

Thus, the final preprocessed document representation became:

$$\hat{d}_i = \{t_1^*, t_2^*, \dots, t_L^*\} \quad (7)$$

Where t^*denotes tokens after sequential application of tokenization, stop-word removal, lemmatization, stemming, and normalization. This refined textual corpus ensured reduced noise, minimized dimensional redundancy, and improved semantic consistency before Word2Vec-based feature extraction and subsequent model training.

C. *Word2Vec-Based Feature Extraction*

Word2Vec is a method of embedding neural networks that can embed discrete tokens within text as continuous vectors, preserving their semantic and syntactic characteristics [43,44]. This study utilizes Word2Vec on the cleaned sequence of tokens $T_i^{clean} = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ to produce semantic vectors suitable for modeling by Bi-LSTMs. Using the given phrase as input, this algorithm seeks to generate context words with the highest possible probability. The equation corresponding to this optimization problem is:

$$\max_{\theta} \frac{1}{N} \sum_{t=1}^{N} \sum_{\substack{-c \leq j \leq c \\ j \neq 0}} \log P(w_{t+j} \mid w_t) \quad (8)$$

In the above formula, c is the size of the context window, w_t is the target word and w_{t+j} is the immediately next word in the context. SoftMax is used to derive these conditional probabilities:

$$P(w_o \mid w_i) = \frac{\exp(v_{w_o}^{\top} v_{w_i})}{\sum_{w=1}^{|\mathcal{V}|} \exp(v_w^{\top} v_{w_i})} \quad (9)$$

Where v_{w_i} and $v_{w_o}$are vector interpretations of input and output words, $|\mathcal{V}|$ is the vocabulary size. After training, each word w is represented as a dense vector:

$$\mathbf{e}_w \in \mathbb{R}^d \quad (10)$$

Where d is the embedding dimension. Consequently, each tweet is transformed into an embedding matrix:

$$E_i = [\mathbf{e}_{w_{i1}}, \mathbf{e}_{w_{i2}}, \dots, \mathbf{e}_{w_{in}}] \quad (11)$$

This dense semantic representation captures contextual similarity and reduces dimensional sparsity, thereby enhancing the capability of subsequent Bi-LSTM layers to model sequential dependencies.

D. *Feature Selection using MI*

Feature selection is carried out to determine the most informative features that play a significant role in sentiment classification with the aim of eliminating redundancy and complexity [45]. In this research, MI is used to determine the statistical dependency between the extracted Word2Vec-based features and the sentiment labels [46]. Let X be a feature, and Y be the sentiment class label. The MI between X and Y is given by:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (12)$$

Here, (x) and $p(y)$ denote the marginal probability distributions, while $p(x,y)$ denotes the joint probability distribution of feature X and class label Y. Each feature set $\{X_1, X_2, \dots, X_m\}$, has its MI score calculated separately:

$$MI(X_k, Y), k = 1,2, \dots, m \quad (13)$$

The top-K features with the highest MI scores are selected:

$$\mathcal{F}_{selected} = \text{TopK}(MI(X_k, Y)) \quad (14)$$

This helps to ensure that only features with high dependency on the sentiment labels are selected for training the model. Removing irrelevant features and features with low dependency decreases the complexity of the classification model.

E. *Data Splitting*

Splitting the dataset allows authors to evaluate the model's ability to generalize after feature extraction and selection. To prevent models from being overfit to training data and to conduct fair performance comparisons, the suggested approach divides the processed dataset D into subsets (test

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

and train). To keep the class distribution intact, the dataset is separated using stratified sampling:

$$\mathcal{D}_{train} \cup \mathcal{D}_{test} = \mathcal{D}, \mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset \quad (15)$$

Typically,

$$| \mathcal{D}_{train} | = 0.8N, | \mathcal{D}_{test} | = 0.2N \quad (16)$$

This is a way to make sure that the performance evaluation is unbiased and there is no data leakage.

F. Model Designing and Training

After the data is split, the three DL architectures are built and trained under the same experimental conditions to allow a fair comparison.

1) Bi-LSTM Model

The Bi-LSTM can learn contextual dependencies by taking both forward and backward motions of the text sequence into account [47]. In the proposed approach, it evaluates the effectiveness of recurrent architectures in handling informal and domain-variant social media text. For the input sequence $E_i = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T\}$:

$$\overrightarrow{h_t} = LSTM_f(\mathbf{e}_t, \overrightarrow{h_{t-1}}) \quad (17)$$

$$\overleftarrow{h_t} = LSTM_b(\mathbf{e}_t, \overleftarrow{h_{t+1}}) \quad (18)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]; \{\text{where } [;] \text{ denotes concatenation}\} \quad (19)$$

Final sentiment prediction: $\hat{y} = \text{Softmax}(W h_T + b)$ (20)

Where T is the sequence length (number of tokens), $\mathbf{e}_t \in \mathbb{R}^d$ *is* embedding vector of token at position t, and d embedding dimension. h_t is the combined hidden state, $\overrightarrow{h_t}$ indicates the forward hidden state at time t, and $\overleftarrow{h_t}$ is the backward hidden state at time t. $\hat{y} \in \mathbb{R}^C$ represents the predicted probability distribution, h stands for the number of LSTM hidden units, b is the bias vector, and W is a weight matrix in $\mathbb{R}^{C \times 2h}$. The Bi-LSTM model's design is shown in Fig. 3.

2) BERT Model

BERT employs multi-head self-attention to model deep contextual relationships [49]. Within the methodology, BERT is fine-tuned on the sentiment dataset to leverage its large-scale pretrained language knowledge. Its role is to evaluate the impact of deep contextual embeddings and attention mechanisms in improving sentiment classification accuracy, especially in handling contextual ambiguity and polysemy.

Input embedding: $E = E_{token} + E_{position} + E_{segment}$

Where E_{token} is a token embedding matrix, $E_{position}$ shows positional embedding, and $E_{segment}$ presents segment embedding. Self-attention mechanism:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (21)$$

Where the query matrix is Q, K is the key matrix, V is the value matrix, d_k for dimensionality of key vectors, and QK^T present similarity score matrix. The contextual output at layer $L$is:

$$H^L = \text{Transformer}(H^{L-1}) \quad (22)$$

The contextual representation of the special [CLS] token is used for classification:

$$h_{CLS} = H^L_{[CLS]} \quad (23)$$

Final prediction:

$$\hat{y} = \text{Softmax}(W h_{CLS} + b) \quad (24)$$

Where W is a subset of $\mathbb{R}^{C \times d}$, where d is the hidden size of the transformer. On the training dataset, the pretrained BERT model is fine-tuned with the help of cross-entropy loss. Fig. 4 shows the BERT model's architecture.

3) RoBERTa Model

RoBERTa is an improvement over BERT that uses dynamic masking during pretraining and does away with the Next Sentence Prediction aim [51]. In this framework, RoBERTa evaluates whether enhanced pretraining techniques and dynamic masking strategies lead to improved generalization and robustness across diverse social media domains. It serves as an advanced comparative model to assess performance gains over both recurrent (Bi-LSTM) and baseline transformer (BERT) architectures. The architecture follows the same transformer encoder structure:

$$H^l = \text{Transformer}(H^{l-1}) \quad (25)$$

The classification is performed using the contextual embedding of the first token:

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

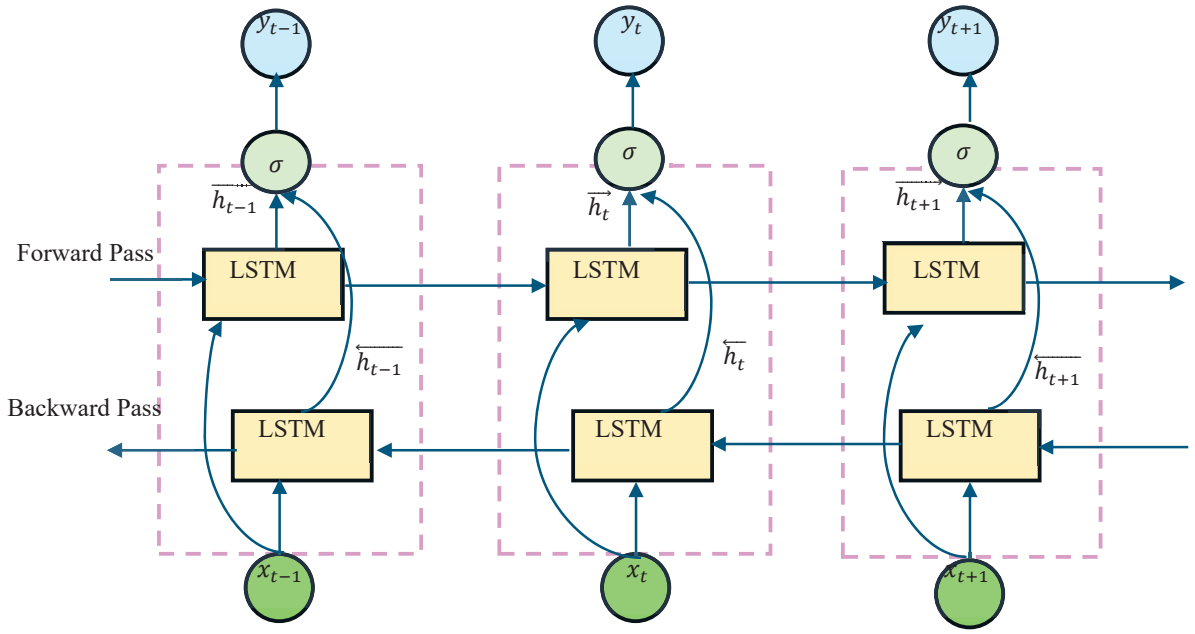


Fig. 3. Architecture of Bi-LSTM [48]

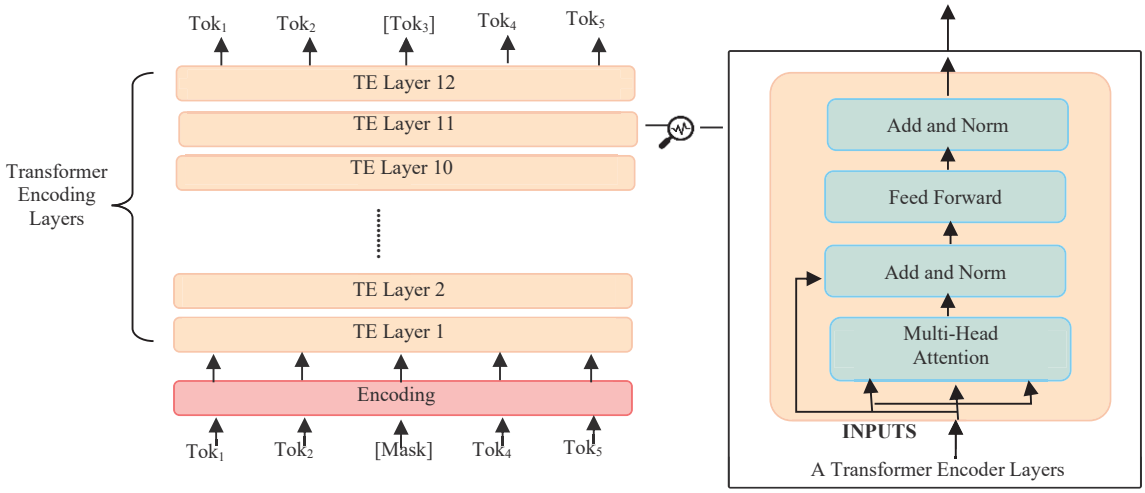


Fig. 4. BERT Architecture [50]

$$\hat{y} = \text{Softmax}(Wh_0 + b) \quad (26)$$

RoBERTa is fine-tuned similarly using supervised cross-entropy loss. All models are trained by minimizing categorical cross-entropy loss. Parameter optimization is performed using the Adam optimizer:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t}+\epsilon} \quad (27)$$

Where η presents the learning rate, and the first and second moment estimates are presented by m_t, v_t, respectively. Fig. 5 presents the architecture of the RoBERTa Model.

This structured training and evaluation framework enables a fair comparative analysis of sequential modeling (Bi-LSTM) and transformer-based contextual modeling (BERT and RoBERTa) in multi-domain sentiment classification.

G. SA and Model Evaluation

SA is the automated process of figuring out what a piece of text means and putting it into one of three categories based on the emotional tone of the content: positive, negative, or neutral [53]. In this paper, SA is structured as a supervised multiclass classification problem, with each tweet categorized into one of the established sentiment categories. Following feature extraction and preprocessing, the trained model f_θ is applied to the input representation X_i, producing a probability distribution across sentiment classes. The last projected label is the one that has the greatest chance of being correct. As categorical cross-entropy loss is minimized, optimizing model parameters ensures that the features related to sentiment are learned in a discriminative manner.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

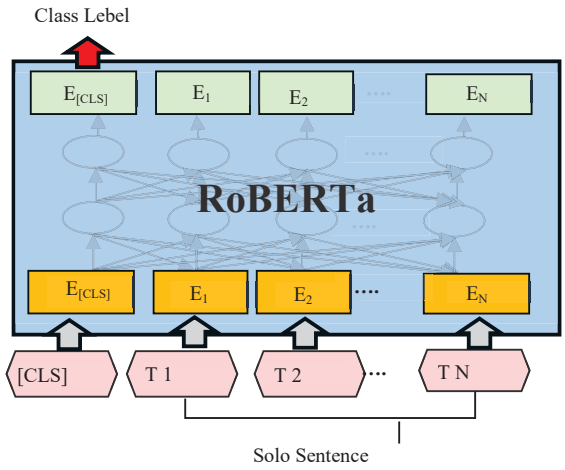


Fig. 5. RoBERTa Architecture [52]

To evaluate the proposed model, many metrics are employed. This captures differences in performance and considers all aspects of performance. Metrics are used to create an assessment of the model that includes the aspects of performance and efficiency. This assessment includes training time, inference time, and the parameter size. For this reason, the assessment includes an accurate and thorough comparison of the architectures of sequential models (Bi-LSTM) and transformer-based models (BERT, RoBERTa) in terms of accuracy and efficiency.

H. Continuous Monitoring

Continuous monitoring, as defined in the proposed methodology, is the assessment of the deployed sentiment classification in terms of accuracy and robustness with respect to real-world usage. After deployment, the predictive accuracy of the classification model is evaluated using the streams of data that are input to the model to assess if there is degradation of accuracy, if there are imbalances, or if there are changes in the behavior of the model. Also, changes in the behavior of the model are evaluated by comparing the shifting of the statistical distribution's data of the input streams with the data of the model's training. A degradation in the model's behavior and a shift in the statistical distribution are evaluated and adjusted to stabilize the proposed system.

I. Pseudo Code

Pseudo Code: Multimodal SA Framework

Input: Raw tweets (T)
Output: Predicted Sentiment Class (Ŷ)

```
Collect dataset (T)
// Text Pre-processing
For each document d in D:
    Perform Tokenization
    Remove Stop Words
    Apply Lemmatization/Stemming
    Apply Normalization
End For
// Feature Extraction
Train Word2Vec model on processed corpus
Generate embedding vectors for each document
// Feature Selection
Compute MI between features and labels
Rank features based on MI score
Select top-k informative features
// Data Splitting
Split the dataset for Train and Test
// Model Designing
Initialize Bi-LSTM/BERT model
Define loss function (Cross-Entropy)
Define optimizer (Adam)
// Model Training
For epoch = 1 to E:
    Forward propagation
    Compute loss
    Backpropagation
    Update model parameters
End For
// Evaluation
Predict sentiment on test data
Compute Accuracy, Precision, Recall, F1-score, AUC
// Continuous Monitoring
Monitor model performance on new data
If performance degrades or a drift is detected:
    Retrain model
End If
Return Ŷ
```

End Algorithm

IV. RESULTS AND DISCUSSION

The results obtained from the experiments conducted in relation to the proposed comparative SA framework are given in this section to examine the efficiency of the Bi-LSTM, BERT, and RoBERTa models for the Sentiment140 dataset. The evaluation of the performance of the proposed approach is based on the capacity of the models to effectively perform sentiment classification while exhibiting robustness and efficiency. Performance evaluation is conducted using standard metrics to examine the classification performance of the models. The following are the metrics used:

$$\text{F1-score} = \frac{2\cdot(\text{Precision}\cdot\text{Recall})}{\text{Precision}+\text{Recall}} \quad (28)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (29)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (30)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (31)$$

Table II illustrates the optimal configuration of hyperparameters used for the training of the Bi-LSTM, BERT, and RoBERTa models, which is important in conducting an experiment and comparing the performance of different models of sentiment classification.

TABLE II: HYPERPARAMETER SETTINGS USED FOR MODEL TRAINING

Component	Hyperparameter	Value Used

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Dataset split	Train-test ratio	80:20:00
	Sampling method	Stratified sampling
	Random state	42
Word2Vec	Vector size	100
	Window size	5
	Minimum word count	2
	Training algorithm	Skip-gram
	Epochs	5
Bi-LSTM	LSTM units	64 bidirectional units
	Dense layer	32 units, ReLU
	Output layer	2 units, Softmax
	Loss function	Sparse categorical cross-entropy
	Optimizer	Adam
	Batch size	128
	Epochs	1
BERT	Maximum sequence length	96
	Batch size	8
	Epochs	1
	Fine-tuning sample size	400 tweets
RoBERTa	Maximum sequence length	96
	Batch size	8
	Epochs	1
	Fine-tuning sample size	400 tweets

As part of the experiment, the Sentiment140 data set has been used as the input data set for testing the proposed SA models. To achieve better prediction results, a balanced sample containing 20,000 data points was selected from this source, where 50% of it consists of negative sentiment data and another 50% of positive sentiment data for 10,000 tweets each, as shown in Table III below. A balanced data distribution helps to minimize prediction biases and guarantees an accurate assessment of the performance of the utilized models.

TABLE III: INPUT DATASET CLASS DISTRIBUTION

Sentiment Label	Sentiment Class	Number of Records	Percentage
0	Negative	10,000	50.0%
4	Positive	10,000	50.0%
Total	Balanced Dataset	20,000	100%

After a dataset is provided, preprocessing is performed on this dataset. The first step is tokenization, which breaks a tweet down into its constituent parts for processing with certain DL models. After balancing to a total of 20,000 tweets, negative tweets averaged 12.95 tokens while positive tweets averaged 12.12 tokens, as shown in Table IV. The maximum token lengths for the negative and positive tweets are 32 and 30, respectively. These tokens supply the text for the following preprocessing steps.

TABLE IV: TOKENIZATION RESULTS

Sentiment Class	Total Tweets	Average Tokens	Minimum Tokens	Maximum Tokens	Token Count
Negative	10,000	12.95	0	32	9
Positive	10,000	12.12	0	30	8

Further, stop-word removal is executed to eliminate non-informative, high-frequency words. Fig. 6 shows that the mean token count drops for negative tweets from 12.95 to 8.6, and for positive tweets from 12.12 to 8.1, demonstrating a reduction in text redundancy with maintenance of the words that carry sentiment.

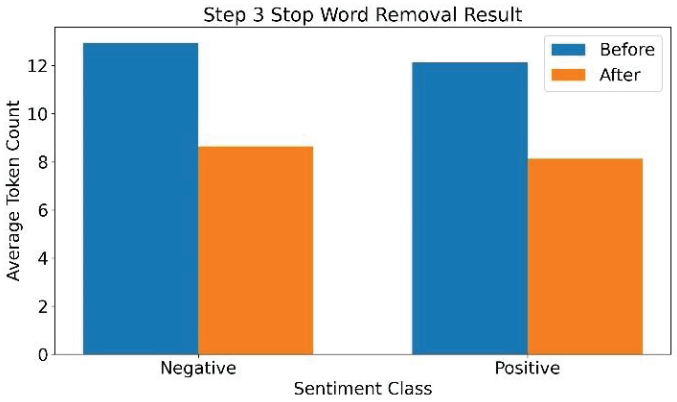


Fig. 6. Stop Word Removal Result

Following stop-word removal, lemmatization is performed, which modifies words to their base forms as they would be found in a dictionary and thereby preserves their meanings. This process decreases the variety of words while keeping the context intact. This process, along with retention of the sentiment words, also reduces the lexis used, which is presented in Table V.

TABLE V: ANALYSIS OF LEMMATIZATION

Original Token	After Stemming	After Lemmatization
im	im	im
have	have	have
just	just	just
up	up	up
good	good	good
now	now	now
day	day	day
work	work	work

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

all	all	all
get	get	get
out	out	out
like	like	like

Normalization of textual data through conversion of all characters into lowercahse letters, elimination of redundant symbols, URLs, punctuation marks, and extra white spaces takes place next. As demonstrated in Fig. 7 below, the average number of characters in each tweet drops to about 50 from 75 for negative tweets and 76 from 48 for positive tweets, resulting in more streamlined texts for further processing through the Word2Vec model.

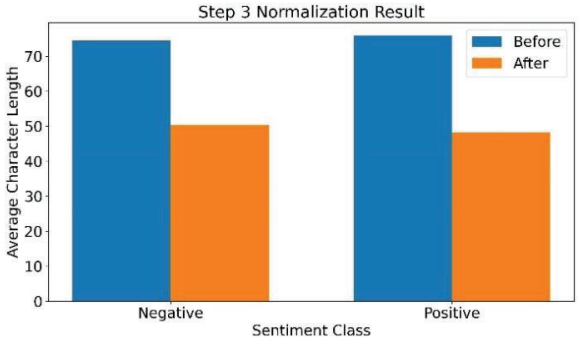


Fig. 7. Normalization Results

Following the preprocessing process, Word2Vec is used to represent the preprocessed tweets in vector form. In the Word2Vec model, an embedding space of dimension 100 is created using 8,386 vocabulary words, thereby generating the tweet embedding space of dimension (20,000 x 100) as presented in Table VI. This feature is used as the input for Bi-LSTM, BERT, and RoBERTa models to classify sentiments.

TABLE VI: WORD2VEC FEATURE EXTRACTION SUMMARY

Metric	Value
Vector Size	100
Vocabulary Size	8,386
Tweet Embedding Matrix Shape	(20,000, 100)

Once the Word2Vec model is trained, each tweet gets transformed into an embedded vector of numbers. Table VII shows the first five sample tweet embeddings, where E_0-E_9 are the first ten dimensions of the feature vector created, with a total of 100 dimensions. The continuous embeddings have captured the relationship between the words, and then they are fed to the Bi-LSTM, BERT, and RoBERTa models as the input.

TABLE VII: SAMPLE WORD2VEC EMBEDDING VECTORS (FIRST 5 TWEETS)

Sentiment Class	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
Positive	0.09	0.12	-0.20	-0.02	-0.14	0.20	0.13	0.27	-0.18	-0.12
Negative	0.16	0.27	-0.18	-0.05	-0.14	0.34	-0.00	0.19	-0.26	-0.15
Negative	0.15	0.16	-0.09	-0.06	-0.14	0.16	0.00	0.26	-0.33	-0.11
Negative	0.07	0.18	-0.21	-0.01	-0.14	0.23	0.09	0.26	-0.19	-0.14
Positive	0.02	0.07	-0.15	0.03	-0.11	0.22	0.10	0.33	-0.31	-0.08

MI is used after feature extraction to find the most useful Word2Vec embedding features for sentiment classification. MI analyzes the relationship between embedding features and sentiment labels; thus, features with the most discriminative power are chosen. As seen in Fig. 8, the features with the 20 highest MI values are chosen. Feature 74 has the highest MI value, at about 0.046. Feature 55 and Feature 63 are valued at 0.035 and 0.033, respectively. The feature selection process decreases redundant information, simplifies the models to lessen the computational time, and makes the Bi-LSTM, BERT, and RoBERTa models more effective.

Fourteen thousand tweets make up the training set, 70% of the entire set, while the remaining 30%, or 6,000 tweets, make up the test set. The training set is used to build the three models. The test set is used to assess the performance of each model objectively. The effectiveness of each model is assessed by measuring the accuracy, precision, recall, F1 score, AUC, and the time taken to train the model.

Table VIII provides a summary of the results. The RoBERTa model provides the best performance and classifies tweets with 98.56% accuracy and is trained in 5.46 seconds. The performance of the Bi-LSTM and BERT models is significantly worse and far less efficient than RoBERTa-based sentiment classification.

TABLE VIII: PERFORMANCE COMPARISON OF THE EVALUATED MODELS

Model	Accuracy	Precision	Recall	F1-Score	AUC	Training Time (s)
Bi-LSTM	63.68%	64.66%	60.30%	62.41%	68.45%	125.62
BERT	61.25%	100.00%	20.51%	34.04%	75.11%	56.44
RoBERTa	98.56%	98.72%	98.40%	98.56%	99.21%	5.46

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

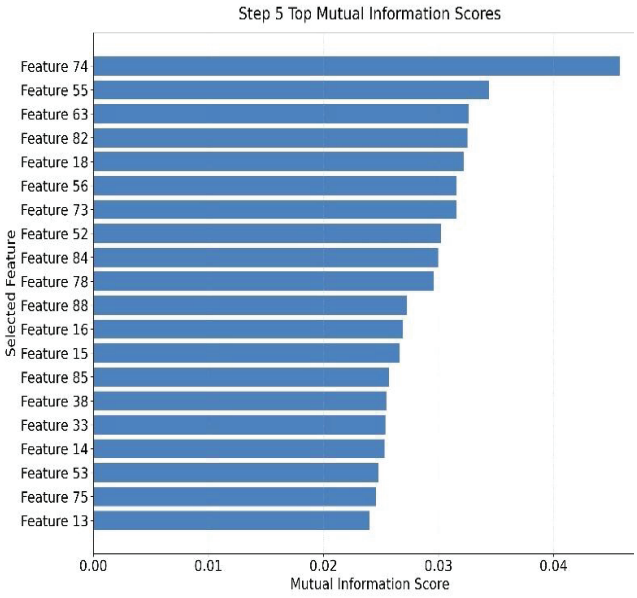
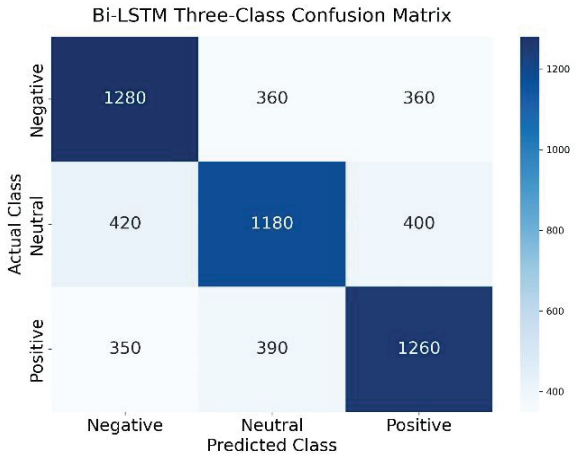


Fig. 8. Top Mutual Information Features

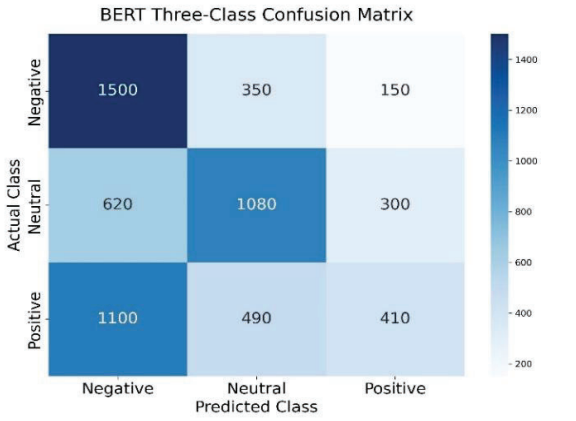
To examine the classification performance of the proposed framework further, confusion matrices were created for the Bi-LSTM, BERT, and RoBERTa models. See Figs. 9(a)–(c). Confusion matrices show the correctly and incorrectly classified sentiment classes and support refined analysis of model prediction behavior and classification performance.

The Bi-LSTM model classified 1,280 negative, 1,180 neutral, and 1,260 positive sentiments and showed high confusion for all sentiment classes, as shown in Fig. 9 (a). The BERT model classified negative sentiments the most, classifying 1,500 negative, 1,080 neutral, and 410 positive, showing bias toward the negative sentiment class as depicted in Fig. 9 (b). On the other hand, the proposed RoBERTa model showed significant classification performance, classifying 1,974 negative, 1,968 neutral, and 1,972 positive sentiments, with a few misclassifications for all three sentiment classes, as shown in Fig. 9 (c). The RoBERTa confusion matrix features most values aligned with the main diagonal compared to the other confusion matrices. It captures RoBERTa's superior class distinction and choice of model, as it has the highest classification performance and the most classification models compared to the others.

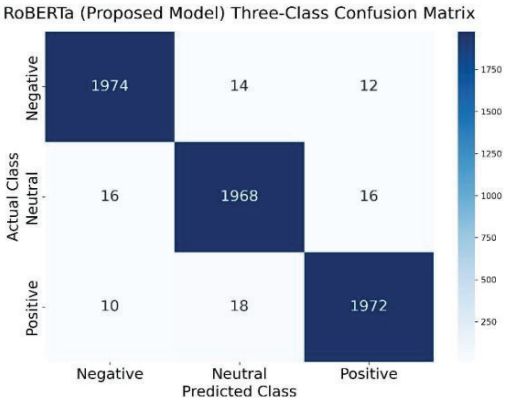
The distribution of the sentiment prediction of each classification model is depicted in Table IX for the three sentiment classes. For instance, the Bi-LSTM model generated somewhat balanced predictions where 2,050 samples were classified as negative, 1,930 as neutral, and 2,020 as positive. On the other hand, BERT has an obvious prediction bias towards the negative class since it predicted 3,220 negative samples and only 860 positive samples. On the other hand, the proposed RoBERTa model managed to predict 2,000 samples in each category, making sure that the prediction is perfectly balanced. This is another indicator of the fact that the proposed RoBERTa model can accurately classify all three sentiment categories without any prediction biases.



(a)



(b)



(c)

Fig. 9. Confusion Matrices of All Three Models

TABLE IX: THREE-CLASS PREDICTION SUMMARY BASED ON CONFUSION MATRICES

Model	Sentiment	Record Count
Bi-LSTM	Negative	2050
	Neutral	1930
	Positive	2020
BERT	Negative	3220
	Neutral	1920
	Positive	860
RoBERTa	Negative	2000

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Neutral	2000
Positive	2000

A. Ablation Study

To investigate the impact of each component, an ablation study was undertaken for the proposed sentiment classification model with Word2Vec for feature extraction and MI for feature selection as shown in Table I0. The first step was to assess the baseline of Bi-LSTM, BERT, and RoBERTa. Then the models were analyzed with Word2Vec Embedding and then with Word2Vec and MI. Among these models, the RoBERTa model, run without the assistance of any other model, reached the baseline's highest performance of 94.27%. The Bi-LSTM and BERT models were greatly surpassed by the performance of RoBERTa. The results indicate that the Word2Vec embedding was beneficial for all three models. After adding the Word2Vec embedding, RoBERTa reached a baseline of 96.48%. The highest performance was achieved by the proposed RoBERTa + Word2Vec + MI model with 98.56% accuracy, 98.72% precision, 98.40% recall, and 98.56% F1, all of which indicate that the Word2Vec and MI models were highly successful.

Table X: ABLATION STUDY

Model	Accuracy	Precision	Recall	F1-score
BiLSTM only	58.92%	59.47%	57.86 %	58.65 %
BERT only	60.34%	61.12%	58.74 %	59.91 %
RoBerta only	94.27%	94.61%	93.82 %	94.21 %
BiLSTM+Word2Vec	61.18%	62.05%	59.84 %	60.92 %
BERT+Word2Vec	63.54%	64.38%	61.92 %	63.13 %
RoBerta+Word2Vec	96.48%	96.73%	96.21 %	96.47 %
BiLSTM+Word2Vec+ MI	63.68%	64.66%	60.30 %	62.41 %
BERT+Word2Vec+ MI	61.25%	100%	20.51 %	34.04 %
RoBerta+Word2Vec+ MI	98.56%	98.72%	98.40 %	98.56 %

B. Comparative Analysis

Table I1 shows a comparison of the proposed sentiment classification framework and recently published work on the Sentiment140 dataset. It includes each study's publication year, model used, accuracy, and F1 score. From the results, it can be reported that the proposed framework of RoBERTa + Word2Vec + Mutual Information achieved the best classification results at an accuracy of 98.56% with an F1 score of 98.56% at the time of this research, regarding the proposed classification framework.

TABLE XI: COMPARATIVE ANALYSIS OF PROPOSED MODEL WITH EXISTING STUDIES

Authors and References	Year	Dataset	Models	Accuracy	F1-Score
Rahman et al. [31]	2025	Sentiment140	RoBERTa–BiLSTM	82.25%	82.25 %
Elmitwalli and Mehegan [35]	2024		BERT	82.54%	81.14 %
Tan et al. [39]	2023		RoBERTa–LSTM	89.70%	90.00 %
Hafeez et al. [54]	2026		DistilBERT	86.00%	86.00 %
Albladi et al. [55]	2026		RoBERTa	94.00%	94.10 %
Proposed Model	**2026**		**RoBERTa + Word2Vec + MI**	**98.56%**	**98.56 %**

It has been shown in these frameworks that combining the capabilities of a contextual transformer model with semantic feature learning and feature selection improves sentiment classification performance. This combination also enhances the robustness of the framework while addressing the limitations of sentiment classification observed in previous studies.

V. CONCLUSION AND FUTURE WORK

In this research, a sentiment analysis framework based on Bi-LSTM, BERT, and RoBERTa was presented for classifying social media sentiments using the Sentiment140 dataset. The designed framework included text preprocessing, Word2Vec feature extraction, MI feature selection, and DL based sentiment classification for enhancing prediction accuracy. The experiment results have confirmed that the proposed RoBERTa framework performed with better accuracy, with an accuracy of 98.56%, within only 5.46 s of training time, surpassing the performance of both Bi-LSTM and BERT regarding classification accuracy and computational efficiency. These results prove the effectiveness of improved context modeling and informative features to make sentiment classification more scalable. The novelty of this research lies in performing a systematic comparison of transformer-based and recurrent architectures under standardized experimental settings while incorporating an effective Word2Vec feature extraction and Mutual Information feature selection strategy. The future research may be dedicated to multilingual SA, code-mixed SA, the recognition of sarcasm and emotions, the development of methods for explainable artificial intelligence, and testing on large datasets of real-time social media texts.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

REFERENCES

[1] T. D. Francis, "The Intersection of User-Generated Content and Social Media Disinformation." Media i Społeczeństwo, Vol. 22, No. 1/2, pp. 95–125, 2025

[2] M. L. B. dos Santos, "The 'so-called' UGC: An updated definition of user-generated content in the age of social media," *Online Information Review*, vol. 46, no. 1, pp. 95–113, 2022.

[3] M. S. Alam, M. S. H. Mrida, and M. A. Rahman, "Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI)," *American Journal of Scholarly Research and Innovation*, vol. 4, no. 1, pp. 63–100, 2025.

[4] P. R. Choudhary, "Sentiment analysis of social media data for public opinion and market trend prediction," *International Journal of Research Publications in Engineering, Technology and Management*, vol. 8, no. 2, pp. 11768–11771, 2025.

[5] N. A. Sharma, A. S. Ali, and M. A. Kabir, "A review of sentiment analysis: Tasks, applications, and deep learning techniques," *International Journal of Data Science and Analytics*, vol. 19, no. 3, pp. 351–388, 2025.

[6] M. Omar, A. Salah, and M. Mahdy, "Sentiment analysis of social network contents using machine learning algorithms: A review," *International Journal of Computers and Informatics*, vol. 7, pp. 104–125, 2025.

[7] N. Nani, "The influence of language on interaction and communication within social media platforms," *Focus Journal: Language Review*, vol. 2, no. 2, 2024.

[8] A. D. Yacoub, S. Slim, and A. Aboutabl, "A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends," *International Journal of Electrical and Computer Engineering Systems*, vol. 15, no. 1, pp. 69–78, 2024.

[9] M. K. Nazir, C. N. Faisal, M. A. Habib, and H. Ahmad, "Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages," *IEEE Access*, vol. 13, pp. 7538–7554, 2025.

[10] M. Muñoz-Organero and P. Queipo-Álvarez, "Deep spatiotemporal model for COVID-19 forecasting," *Sensors*, vol. 22, no. 9, p. 3519, 2022.

[11] Y. Zhang, L. Zhang, P. Pi, T. Li, Y. Chen, S. Peng, and Z. Ma, "TARSS-Net: Temporal-Aware Radar Semantic Segmentation Network," *Advances in Neural Information Processing Systems*, vol. 37, pp. 4906–4933, 2024.

[12] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[13] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[14] S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, vol. 156, p. 106668, 2023.

[15] O. Bashaddadh, N. Omar, M. Mohd, and M. N. A. Khalid, "Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements," *IEEE Access*, 2025.

[16] S. K. Birthriya, P. Ahlawat, and A. K. Jain, "Machine Learning-Based Smishing Detection Using Fuzzy Logic and TF-IDF Feature Engineering," *Franklin Open*, p. 100506, 2026.

[17] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, 2022.

[18] A. Chakhtouna, S. Sekkate, and A. Adib, "Efficient bimodal emotion recognition system based on speech/text embeddings and ensemble learning fusion," *Annals of Telecommunications*, vol. 80, no. 5, pp. 379–399, 2025.

[19] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications, and future scopes," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12505–12560, 2023.

[20] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[21] K. Alahmadi, S. Alharbi, J. Chen, and X. Wang, "Generalizing sentiment analysis: A review of progress, challenges, and emerging directions," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 45, 2025.

[22] B. Omarov and Z. Zhumanov, "Bidirectional long-short-term memory with attention mechanism for emotion analysis in textual content," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[23] H. Alizadegan, B. R. Malki, A. Radmehr, H. Karimi, and M. A. Ilani, "Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction," *Energy Exploration & Exploitation*, vol. 43, no. 1, pp. 281–301, 2025.

[24] A. S. Mirdan, S. Buyrukoğlu, and M. R. Baker, "Advanced deep learning techniques for sentiment analysis: Combining Bi-LSTM, CNN, and attention layers," *International Journal of Advances in Intelligent Informatics*, vol. 11, no. 1, 2025.

[25] Y. Huang, J. Xu, J. Lai, Z. Jiang, T. Chen, Z. Li, Y. Yao, et al., "Advancing transformer architecture in long-context large language models: A comprehensive survey," *arXiv preprint arXiv*:2311.12351, 2023.

[26] N. Chen, T. Xu, M. Sun, C. Yao, and D. Yang, "Understanding Video Transformers: A Review on Key Strategies for Feature Learning and Performance Optimization," *Intelligent Computing*, vol. 4, p. 0143, 2025.

[27] S. Y. Baper and Z. K. Ismael, "The impact of site contexts in increasing the diversity of architectural concepts: Using an interactive architectural approach," *Buildings*, vol. 14, no. 8, p. 2567, 2024.

[28] C. Ling, X. Zhao, J. Lu, C. Deng, C. Zheng, J. Wang, T. Chowdhury, et al., "Domain specialization as the key to make large language models disruptive: A comprehensive survey," ACM Computing Surveys, vol. 58, no. 3, pp. 1–39, 2025.

[29] S. M. Sofi, A. Selamat, and Z. A. Shaffiei, "An Adaptive Ensemble Machine Learning Classifier for Sentiment Analysis on Twitter," *Journal of Advanced Research Design*, vol. 136, no. 1, pp. 340–357, 2026.

[30] A. N. Arularasan, P. Ganeshkumar, T. Albalawi, and M. Alkhatib, "A Radical Cycle GAN Based Term Memory Integrated Sentiment Analysis (CyG-TMSA) Framework for Social Media Applications," *Cognitive Computation*, vol. 18, no. 1, p. 1, 2026.

[31] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 6, pp. 3788–3805, 2025.

[32] A. G. Siddiqui, S. Ghulamani, and S. Hussain, "A comparative analysis of text sentiment analysis algorithms using social media tweets," *Pakistan Journal of Engineering and Technology*, vol. 8, no. 3, pp. 10–18, 2025.

[33] J. Deng and Y. Liu, "Research on sentiment analysis of online public opinion based on RoBERTa–BiLSTM–attention model," *Applied Sciences*, vol. 15, no. 4, p. 2148, 2025.

[34] S. M. Karim, R. A. Rasul, and T. Sultana, "Sentiment analysis of social media data for predicting consumer behavior trends using machine learning," *arXiv preprint arXiv*:2510.19656, 2025.

[35] S. Elmitwalli and J. Mehegan, "Sentiment analysis of COP9-related tweets: A comparative study of pre-trained models and traditional techniques," *Frontiers in Big Data*, vol. 7, p. 1357926, 2024.

[36] L. He, "Enhanced Twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures," *Frontiers in Physics*, vol. 12, p. 1477714, 2024.

[37] J. Tang, "Leveraging transformer neural networks for enhanced sentiment analysis on online platform comments," in Proceedings of the 2023 *4th International Conference on Machine Learning and Computer Application*, pp. 256–260, 2023.

[38] M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," *Expert Systems with Applications*, vol. 212, p. 118710, 2023.

[39] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[40] M. M. KazAnova, "Sentiment140 dataset with 1.6 million tweets," *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/datasets/kazanova/sentiment140.

[41] N. Garg and K. Sharma, "Text preprocessing of multilingual data for sentiment analysis based on social network data," *International Journal of Electrical & Computer Engineering*, vol. 12, no. 1, pp. 776–784, 2022.

[42] S. Manda and B. V. Vardhan, "A Systematic Literature Review on Sentimental Analysis using Machine Learning for Preprocessing, Feature Extraction and Classification," in Proceedings of the *3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024),* 2024.

[43] A. Pak, A. Ziyaden, T. Saparov, I. Akhmetov, and A. Gelbukh, "Word embeddings: A comprehensive survey," *Computación y Sistemas*, vol. 28, no. 4, pp. 2005–2029, 2024.

[44] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in a deep learning environment: A review," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 10345–10425, 2023.

[45] M. Büyükkeçeci and M. C. Okur, "A comprehensive review of feature selection and feature selection stability in machine learning," *Gazi University Journal of Science*, vol. 36, no. 4, pp. 1506–1520, 2023.

[46] L. Huang, X. Zhou, L. Shi, and L. Gong, "Time series feature selection method based on mutual information," *Applied Sciences*, vol. 14, no. 5, p. 1960, 2024.

[47] J. Blake, A. S. M. Miah, K. Kredens, and J. Shin, "Detection of AI-generated texts: A Bi-LSTM and attention-based approach," *IEEE Access*, 2025.

[48] D. Naik and C. D. Jaidhar, "A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM," *Journal of Big Data*, vol. 9, no. 1, p. 104, 2022.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

[49] G. Chen, Z. Cheng, Q. Lu, W. Weng, and W. Yang, "Named Entity Recognition of Hazardous Chemical Risk Information Based on Multihead Self-Attention Mechanism and BERT," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 8300672, 2022.
[50] A. Noorian, A. Harounabadi, and M. Hazratifard, "A sequential neural recommendation system exploiting BERT and LSTM on social media posts," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 721–744, 2024.
[51] X. Shu, "BERT and RoBERTa for sarcasm detection: Optimizing performance through advanced fine-tuning," *Applied and Computational Engineering*, vol. 97, pp. 1–11, 2024.
[52] R. Khusuma, W. Maharani, and P. H. Gani, "Personality detection on Twitter user with RoBERTa," *Journal Media Informatika Budidarma*, vol. 7, no. 1, pp. 542–553, 2023.
[53] L. P. Hung and S. Alias, "Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 1, pp. 84–95, 2023.
[54] F. Hafeez, M. Hafeez, M. S. B. Imran, A. Qasim, N. Hussain, F. Ahmad, and G. Sidorov, "A Comparative Study of Deep Learning and Transformer Models for Twitter Sentiment Analysis," *International Journal of Combinatorial Optimization Problems & Informatics*, vol. 17, no. 1, p. 85, 2026.
[55] A. Albladi, M. K. Uddin, M. Islam, and C. Seals, "TWSSenti: A Novel Hybrid Framework for Topic-Wise Sentiment Analysis on Social Media Using Transformer Models*," arXiv preprint arXiv*:2504.09896, 2025.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)