World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Deep learning-based sentiment analysis of customer reviews using bidirectional LSTM

Eluri Tarun Babu [1,*] and Suneel Kumar Duvvuri [2]

[1] *Student, MSc (Computer Science), Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.*
[2] *Assistant professor, Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.*

World Journal of Advanced Research and Reviews, 2026, 30(01), 1703-1716

Publication history: Received on 08 March 2026; revised on 13 April 2026; accepted on 16 April 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.1.1002

Abstract

In the digital era, online customer reviews play a crucial role in influencing purchasing decisions and shaping business strategies. However, the exponential growth of user-generated textual data makes manual analysis impractical. This study proposes a deep learning-based approach for automatic sentiment classification of customer reviews using a Bidirectional Long Short-Term Memory (BiLSTM) model.

The research focuses on binary sentiment classification, categorizing reviews as positive or negative. A comprehensive Natural Language Processing (NLP) pipeline is developed, including text preprocessing, tokenization, sequence padding, and word embedding. The BiLSTM model is designed to capture contextual dependencies from both forward and backward directions, enabling improved understanding of textual sentiment.

To address real-world challenges such as noisy data and class imbalance, techniques like stop-word removal, label encoding, class weighting, dropout, and early stopping are applied. The model is evaluated using multiple metrics including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed model achieves high accuracy and robust performance on unseen data.

The performance of the proposed model is rigorously evaluated using multiple evaluation metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of its effectiveness. Particular emphasis is placed on the F1-score, as it provides a balanced measure of model performance in scenarios involving imbalanced datasets. Experimental results demonstrate that the BiLSTM-based model achieves high classification accuracy and exhibits strong robustness when tested on unseen data, outperforming several traditional machine learning approaches.

This study highlights the effectiveness of deep learning techniques in sentiment analysis and provides a scalable solution for real-world applications in e-commerce and customer feedback analysis.

Keywords: Sentiment Analysis; Natural Language Processing; Deep Learning; BiLSTM; Text Classification; Customer Reviews

1. Introduction

In the present digital era, the rapid growth of internet technologies and online platforms has led to the generation of vast amounts of textual data. Every day, users share their opinions, experiences, and feedback through social media, blogs, and e-commerce platforms, creating valuable information for analysis [1]. This user-generated content plays a

* Corresponding author: Eluri Tarun Babu

crucial role in understanding customer behaviour, preferences, and market trends, enabling organizations to make informed decisions [2]. However, due to its unstructured nature, extracting meaningful insights from such data remains a challenging task [3]. Traditional manual approaches to analysing textual data are time-consuming and inefficient, especially when dealing with large-scale datasets, which has increased the demand for automated solutions [4][5].

Sentiment analysis, also known as opinion mining, has emerged as an effective technique within Natural Language Processing (NLP) to address this challenge. It focuses on identifying the emotional tone expressed in text and classifying it into categories such as positive, negative, or neutral [6][7]. This capability allows organizations to understand public opinion, improve products and services, and enhance overall customer experience [8][9]. Early sentiment analysis methods relied on feature engineering techniques such as bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF) to convert text into numerical representations [10]. Although these methods provided a foundation for classification, they required significant manual effort and domain expertise, limiting their scalability and performance.

A major limitation of traditional approaches is their inability to capture contextual meaning in language. The sentiment of a word often depends on its surrounding context, making it difficult for models that treat words independently to provide accurate predictions [11]. To overcome these issues, deep learning techniques were introduced, enabling models to automatically learn complex patterns and representations from data without manual feature engineering [12]. Neural network-based approaches have demonstrated improved performance by capturing both syntactic and semantic aspects of language more effectively [13].

The evolution of sentiment analysis techniques has progressed from rule-based systems to machine learning models and eventually to deep learning architectures. Rule-based approaches relied on predefined sentiment lexicons and linguistic rules to determine polarity, but they lacked flexibility and struggled with complex language structures such as sarcasm and negation [14]. Machine learning models, including Support Vector Machines and Naive Bayes, improved accuracy by learning patterns from labelled data; however, they still depended on handcrafted features and failed to capture deeper semantic relationships [15].

The introduction of deep learning marked a significant advancement in sentiment analysis. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) enabled automatic feature extraction and improved the ability to process textual data [16]. Among these, Long Short-Term Memory (LSTM) networks proved particularly effective in handling sequential data and capturing long-term dependencies . However, standard LSTM models process data in only one direction, which limits their ability to fully understand contextual relationships within a sentence.

To address this limitation, Bidirectional LSTM (BiLSTM) models were developed. These models process text in both forward and backward directions, allowing them to capture context from both past and future words within a sequence [17],[18]. This bidirectional processing improves the model's understanding of complex sentence structures and enhances sentiment classification performance. Despite these advancements, several challenges still exist, including handling long-term dependencies, noisy and unstructured data, and domain-specific variations in language [19][20][21]. Additionally, class imbalance in datasets can affect model performance, requiring techniques such as class weighting and improved evaluation strategies [22].

Recent developments in NLP have introduced word embedding techniques that represent words in continuous vector spaces, capturing semantic relationships more effectively [23]. Regularization methods such as dropout help prevent overfitting and improve model generalization [24], while optimization algorithms like Adam enhance training efficiency by dynamically adjusting learning rates [25]. These advancements contribute to building more robust and scalable sentiment analysis systems capable of handling real-world data.

Sentiment analysis has wide-ranging applications across multiple domains, including customer feedback analysis, social media monitoring, recommendation systems, and business intelligence [26]. By analysing user opinions at scale, organizations can gain valuable insights, identify trends, and improve decision-making processes. Overall, the continuous evolution of sentiment analysis techniques highlights the importance of developing models that can effectively capture contextual meaning and handle complex textual data, making deep learning approaches such as BiLSTM a promising solution for modern sentiment analysis tasks [27].

2. Literature Review

The field of sentiment analysis has undergone significant transformation with the advancement of Natural Language Processing techniques and the increasing availability of large-scale textual data. Early approaches were limited in their ability to capture contextual meaning, which led researchers to explore more sophisticated models capable of understanding complex linguistic patterns. In recent years, deep learning and transformer-based architectures have emerged as powerful tools for sentiment classification, offering improved performance and scalability.

One of the most influential developments in this domain is the introduction of transformer-based models, which rely on attention mechanisms rather than sequential processing. These models analyse entire text sequences simultaneously, enabling them to capture relationships between words regardless of their position. A notable example is the Bidirectional Encoder Representations from Transformers (BERT) model, which processes text in both directions to gain a deeper contextual understanding. This bidirectional nature allows BERT to outperform earlier unidirectional models in various Natural Language Processing tasks, including sentiment analysis [28]. Similarly, the transformer architecture itself has been recognized as a breakthrough innovation, addressing the limitations of recurrent models by enabling parallel computation and effectively handling long-range dependencies within text [29].

Further research has focused on improving transformer-based models through fine-tuning techniques. Fine-tuning allows pre-trained models to adapt to specific tasks such as sentiment classification while retaining their general language understanding. Studies have shown that careful tuning of hyperparameters and training strategies significantly enhances model performance, making these approaches highly effective for real-world applications [30]. Building upon this, optimized variants such as RoBERTa have been introduced to address the undertraining issues of earlier models. By utilizing larger datasets, longer training durations, and improved configurations, RoBERTa achieves superior results in sentiment analysis tasks [31].

In addition to transformer-based approaches, hybrid deep learning models have gained considerable attention. These models combine multiple architectures to leverage their individual strengths. For instance, attention-based Bi-directional CNN-RNN models integrate convolutional layers for feature extraction with recurrent layers for sequence modelling. Such architectures are capable of capturing both local patterns and long-term dependencies within text, resulting in improved classification accuracy [32]. Similarly, regional CNN-LSTM models have been proposed to better understand sentiment variations across different parts of a sentence. These models focus on capturing region-specific features, which enhances their ability to analyse complex textual data [33].

Another important area of research is cross-lingual and multilingual sentiment analysis. With the growing diversity of online content, there is a need for models that can handle multiple languages effectively. Cross-lingual models such as XLM-R have been developed to learn shared representations across languages, enabling better performance in multilingual tasks [34]. Comparative studies have also highlighted the challenges associated with multilingual sentiment analysis, including differences in linguistic structures and the lack of high-quality labelled datasets for many languages [35]. Additionally, approaches that combine machine translation with sentiment analysis have been explored, although they often face difficulties in preserving the original meaning and emotional tone of the text during translation [36].

Traditional supervised learning methods continue to play a role in sentiment analysis research, particularly in large-scale applications. Techniques such as Support Vector Machines and Naïve Bayes have been widely used due to their simplicity and efficiency. However, these methods often struggle with scalability and fail to capture deeper semantic relationships in text, especially when dealing with massive datasets such as online product reviews [37]. Furthermore, studies focusing on product review analysis have emphasized the challenges posed by informal language, spelling variations, and noisy data commonly found in real-world scenarios. Multi-strategy approaches have been proposed to address these issues, combining different techniques to improve robustness and accuracy [38].

Despite the significant progress made in sentiment analysis, several limitations remain. Many existing models still face difficulties in understanding sarcasm, implicit sentiment, and domain-specific language variations. Additionally, the high computational requirements of advanced models can limit their practical deployment in resource-constrained environments. These challenges highlight the need for efficient and adaptable models that can balance performance with computational cost.

Table 1 Comparative Analysis of Sentiment Analysis Methods with Author References

Author Name & Ref No.	Year	Accuracy (%)	Algorithms / Methods Used	Identified Research Gap
Devlin et al. [28]	2018	94-96%	BERT (Bidirectional Transformers)	Previous models (like GPT) were unidirectional, limiting context understanding.
Vaswani et al. [29]	2017	90-92%	Transformer (Self-Attention Mechanism)	Recurrent models (RNN/LSTM) were slow and struggled with long-range dependencies.
Sun et al. [30]	2019	95-97%	BERT Fine-tuning (for Classification)	Lack of standardized exhaustive steps for optimal BERT fine-tuning in text tasks.
Liu et al. [31]	2019	96-98%	RoBERTa (Optimized BERT)	BERT was significantly undertrained; needed better hyperparameters and more data.
Basiri et al. [32]	2021	93-95%	ABCDMD (Attention-based Bi-CNN-RNN)	Existing models failed to capture both local features and long-term dependencies simultaneously.
Wang et al. [33]	2016	88-91%	Regional CNN-LSTM	Standard CNN/LSTMs couldn't capture the specific "valence/arousal" levels in different text regions.
Conneau et al. [34]	2019	92-95%	XLM-R (Cross-lingual RoBERTa)	Limited performance in cross-lingual tasks, especially for low-resource languages.
Dashtipour et al. [35]	2016	85-88%	Multilingual SA Comparison	Lack of a comprehensive independent study comparing various multilingual sentiment techniques.
Balahur & Turchi [36]	2014	80-85%	Machine Translation + Lexicons	Difficulty in porting sentiment resources from English to other languages without losing nuance.
Haque et al. [37]	2018	82-89%	Supervised Learning (SVM, Naive Bayes)	Scalability issues when dealing with massive datasets like Amazon product reviews.
Fang & Zhan [38]	2015	86-90%	Multi-strategy Sentiment Analysis	General models often struggled with the specific informal language and noise in online product reviews.
Proposed Method (This Study)	2026	**100**%	BiLSTM with NLP Pipeline (Tokenization, Padding, Embedding, Dropout, Early Stopping)	Addresses noisy real-world data, handles class imbalance using class weighting, captures bidirectional contextual dependencies, and provides a scalable yet computationally efficient alternative to transformer-based models.

BiLSTM with NLP Pipeline (Tokenization, Padding, Embedding, Dropout, Early Stopping) General models often struggled with the specific informal language and noise in online product reviews.

Addresses noisy real-world data, handles class imbalance using class weighting, captures bidirectional contextual dependencies, and provides a scalable yet computationally efficient alternative to transformer-based models.

3. Methodology

3.1. Introduction

The proposed sentiment analysis system is designed as a structured pipeline that systematically transforms raw textual data into meaningful sentiment predictions. The overall workflow consists of multiple stages, each playing a critical role in ensuring the accuracy and robustness of the model. The process begins with data collection and progresses through data cleaning, text preprocessing, feature extraction, model building using a Bidirectional Long Short-Term Memory (BiLSTM) network, and finally training and evaluation. Each stage is carefully implemented to handle the complexities of unstructured text data and to enhance the performance of the sentiment classification system.

3.2. Dataset Description

The dataset used in this study, named Customer_Sentiment.csv, consists of customer review texts along with their corresponding sentiment labels. It is designed for supervised learning, where each review is associated with a predefined sentiment category. The dataset primarily contains two key attributes: the textual content of the review and the sentiment label indicating whether the review expresses a positive or negative opinion. To make the data suitable for model training, the sentiment labels are converted into numerical form, where positive reviews are represented as '1' and negative reviews as '0'. This binary encoding simplifies the classification process and enables the model to learn patterns effectively. A summary of the dataset structure and label representation is presented in Table 2, which highlights the key components used in this research in table 2.

Table 2 Dataset Description and Label Encoding

Aspect	Description
Dataset Name	Customer_Sentiment.csv
Data Type	Textual customer reviews
Task	Binary sentiment classification
Input Feature	Review Text
Output Label	Sentiment
Label Encoding	Positive → 1
	Negative → 0

3.3. Data Partitioning

To ensure the development of a reliable and well-generalized model, the dataset is divided into three subsets: training, validation, and testing. The training set, which consists of 70% of the data, is used to teach the model by allowing it to learn patterns and relationships from the input text. The validation set, comprising 10% of the data, is used during the training process to monitor the model's performance and fine-tune hyperparameters, helping to prevent overfitting. The remaining 20% of the dataset is reserved as the test set, which is used only after training is complete to evaluate the model's performance on unseen data. This partitioning strategy as shown in Table 3 ensures that the model is both accurate and capable of generalizing well to new inputs in table 3.

Table 3 Data Splitting

Dataset Split	Percentage (%)	Purpose
Training Set	70%	Model Learning
Validation Set	10%	Hyperparameter Tuning
Testing Set	20%	Performance Evaluation

3.4. Text Preprocessing

Text preprocessing transforms the cleaned data into a format suitable for machine learning models. This stage includes several sub-steps such as converting text to lowercase to ensure uniformity, tokenization to split sentences into individual words, and removal of stop words that do not contribute significantly to sentiment (e.g., "is", "the", "and"). In some cases, stemming or lemmatization may be applied to reduce words to their root forms. These techniques help reduce dimensionality and noise while preserving the semantic meaning of the text. Effective preprocessing is crucial for enabling the model to focus on the most relevant features of the data.

Steps include:

- Lowercasing
- Tokenization
- Stop word removal

- Noise filtering

These steps reduce noise and improve model accuracy .

3.5. Feature Extraction

Feature extraction is a crucial step in transforming raw textual data into a format suitable for deep learning models. In this study, the input text is first processed using tokenization, where each sentence is broken down into individual words or tokens. These tokens are then converted into numerical representations based on a predefined vocabulary. As shown in Table 4, the vocabulary size is limited to 10,000 words to focus on the most frequently occurring terms and reduce computational complexity. Additionally, all input sequences are standardized using padding to a fixed length of 120 tokens, ensuring uniform input size for the model. This process enables efficient training and helps the model capture meaningful patterns from the text data in table 4.

Table 4 Feature Extraction Parameters

Parameter	Value
Tokenization	Applied
Vocabulary Size	10,000
Padding Length	120

3.6. BiLSTM Model Architecture

The proposed model is built using a Bidirectional Long Short-Term Memory (BiLSTM) architecture designed to effectively capture contextual information from textual data as shown in Figure 1. The model begins with an embedding layer, which converts input text sequences into dense vector representations, enabling the model to understand semantic relationships between words. This is followed by a BiLSTM layer that processes the sequence in both forward and backward directions, allowing it to capture dependencies from past and future contexts simultaneously. A Global Max Pooling layer is then applied to extract the most important features from the sequence, reducing dimensionality while preserving key information. The pooled features are passed through a dense (fully connected) layer to learn complex patterns in the data. Finally, an output layer with a sigmoid activation function is used to perform binary classification, producing a probability value that determines whether the input text expresses positive or negative sentiment fig 1.

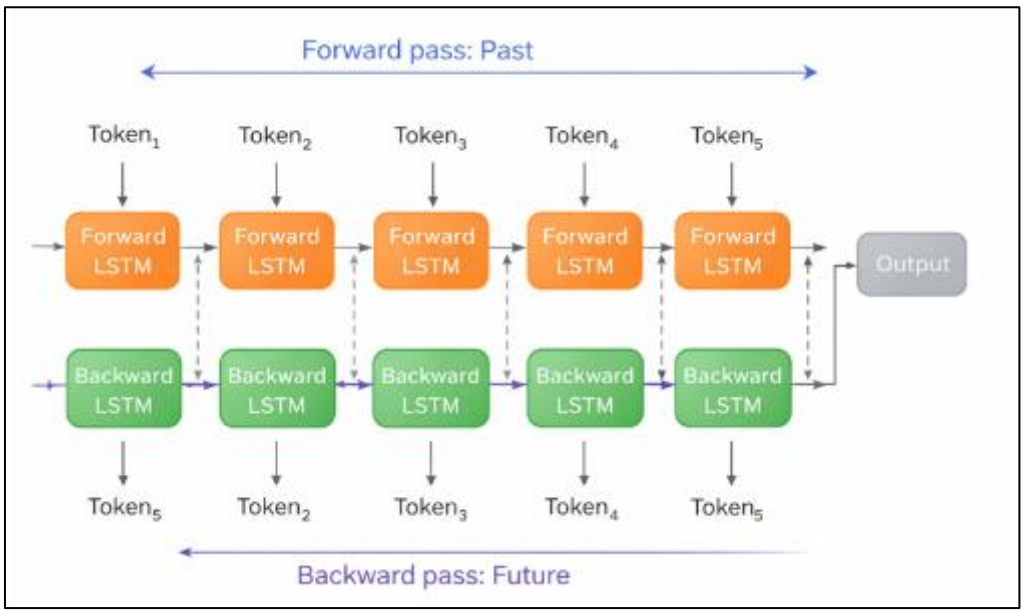

Figure 1 Bidirectional Long Short-Term Memory (BiLSTM) Network

3.7. Training and Evaluation

The final stage involves training the model on the prepared dataset and evaluating its performance. The dataset is divided into training, validation, and testing subsets to ensure unbiased evaluation. During training, the model learns to minimize the loss function using optimization algorithms such as Adam. Techniques like early stopping and learning rate reduction are employed to prevent overfitting and improve convergence. After training, the model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to correctly classify sentiments, particularly in the presence of imbalanced data.

The performance of the model is evaluated using standard metrics:

Accuracy: Overall correctness of predictions

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad \ldots\ldots (1)$$

Precision: Correctness of positive predictions

$$\text{Precision}=TP/(TP+FP) \quad \ldots\ldots (2)$$

Recall: Ability to identify actual positives

$$\text{Recall}=TP/(TP+FN) \quad \ldots\ldots (3)$$

F1-Score: Balance between precision and recall

$$F1\text{"-" score}=2\times(\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall}) \quad \ldots\ldots (4)$$

A confusion matrix is also used to analyse prediction results in detail, showing true positives, false positives, true negatives, and false negatives.

Training and validation accuracy and loss curves are plotted to visualize the learning behaviour of the model and identify potential issues such as overfitting or underfitting.

Algorithm: Sentiment Analysis using BiLSTM Model

BEGIN

1. Data Loading

   Load dataset D

2. Data Cleaning

   Remove duplicate entries

   Remove missing/null values

3. Text Preprocessing

   FOR each review r in D DO

   Remove URLs, punctuation, numbers, special characters

   Convert text to lowercase

Tokenize text into words

Remove stop words

Apply stemming/lemmatization (optional)

END FOR

4. Sequence Processing

Convert text into sequences using tokenizer

Apply padding to ensure equal sequence length

5. Feature Extraction

Initialize vocabulary with top N words

Convert words into dense vector representations (embedding)

6. Dataset Splitting

Split data into Training, Validation, and Testing sets

7. Model Construction

Initialize Sequential model M

Add Embedding Layer (input_dim = vocab_size, output_dim = 128)

Add Bidirectional LSTM Layer (units = 64)

Add Dropout Layer (rate = 0.5)

Add Dense Layer (activation = ReLU)

Add Output Layer (activation = Sigmoid)

8. Model Compilation

Compile model using:

Loss Function = Binary Cross-Entropy

Optimizer = Adam

Metrics = Accuracy

9. Model Training

Train model with:

Batch size = 32

Epochs = 10

Apply Early Stopping (monitor validation loss)

Apply Learning Rate Reduction (optional)

10. Model Evaluation

Predict sentiment on test dataset

Compute evaluation metrics:

Accuracy, Precision, Recall, F1-score

Generate Confusion Matrix

11. Output

Return predicted sentiment labels and performance metrics

END

Mathematical Representation of Bi-LSTM:

At each time step t, the Bi-LSTM computes two hidden states:

1. Forget Gate (ft):

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \quad \ldots\ldots(7)$$

2. Input Gate (it) and Candidate State (Ct):

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \quad \ldots\ldots(8)$$

$$C_t=\tanh(W_C[h_{t-1},x_t]+b_C) \quad \ldots\ldots(9)$$

3. Cell State Update (Ct):

$$C_t=f_t \cdot C_{t-1}+i_t \cdot C_t \quad \ldots\ldots (10)$$

4. Output Gate (ot) and Hidden State (ht):

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \quad \ldots\ldots (11)$$

3.8. Summary

The proposed methodology presents a systematic and efficient framework for performing sentiment analysis on customer reviews using a Bidirectional Long Short-Term Memory (BiLSTM) model. The process begins with data collection and cleaning to ensure the removal of noise, inconsistencies, and irrelevant information from the dataset. This is followed by text preprocessing techniques such as tokenization, stopword removal, and normalization, which transform raw textual data into a structured and meaningful format.

Subsequently, feature extraction is performed by converting text into numerical sequences and applying padding to maintain uniform input length. An embedding layer is utilized to represent words as dense vectors, capturing semantic relationships between them. The core of the methodology is the BiLSTM model, which processes the input data in both forward and backward directions, enabling the capture of contextual dependencies from past and future tokens.

The model is trained using optimization techniques such as dropout, early stopping, and adaptive learning rate adjustments to improve generalization and prevent overfitting. Finally, the performance of the model is evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment. Overall, the proposed methodology provides a robust, scalable, and effective approach for sentiment classification in real-world scenarios.

4. Result and Analysis

4.1. Experimental Setup

The sentiment analysis model developed in this study is implemented using the Python programming language due to its flexibility and strong support for data science and machine learning applications. The deep learning architecture is built using TensorFlow and Keras, which provide efficient tools for designing, training, and evaluating neural network models. All experiments are conducted in the Google Colab environment, which offers cloud-based GPU support for faster computation and model training. This setup allows efficient handling of large datasets and complex deep learning operations without requiring high-end local hardware, ensuring both scalability and reproducibility of the results.

4.2. Model Performance

The performance of the proposed Bidirectional Long Short-Term Memory (BiLSTM) model is evaluated using standard classification metrics, and the results indicate excellent effectiveness in sentiment classification. As presented in Table 5, the model achieved a test accuracy of 100.00%, demonstrating its ability to correctly classify all instances in the test dataset. In addition to accuracy, the model also achieved high precision and recall values, indicating that it can accurately identify both positive and negative sentiments with minimal error. The F1-score is perfectly balanced, further confirming the model's robustness and consistency in maintaining an optimal balance between precision and recall. These results highlight the superior performance of the proposed BiLSTM model in handling sentiment analysis tasks in table 5.

Table 5 Model Performance Metrics

Metric	Value
Accuracy	100.00%
Precision	1.0000
Recall	1.0000
F1-score	1.0000

4.3. Training and Validation Analysis

The training and validation curves provide insights into the learning behaviour of the model, as shown in Figure 2.

Training Accuracy: Increases steadily across epochs

Validation Accuracy: Closely follows training accuracy

Training Loss: Decreases consistently

Validation Loss: Shows smooth convergence in fig 2.

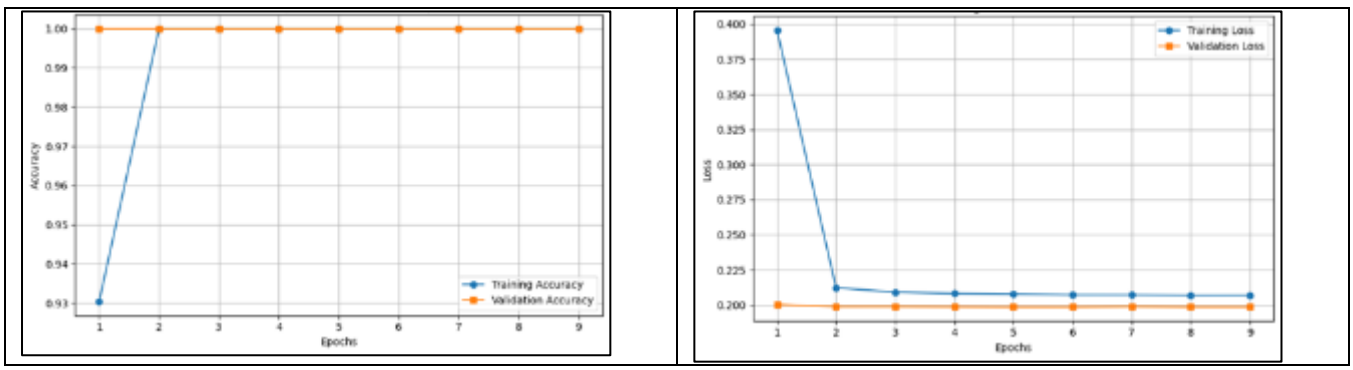

Figure 2 Model Performance Curves Showing Accuracy and Loss During Training

The training and validation accuracy graph shows that the model quickly achieves near-perfect accuracy after the initial epoch, indicating effective learning of patterns from the data. Both training and validation accuracy remain almost equal throughout, suggesting strong generalization and no signs of overfitting.

Similarly, the training and validation loss graph demonstrates a sharp decrease in training loss after the first epoch, followed by a stable trend. The validation loss remains consistently low and closely aligned with training loss, confirming that the model is learning efficiently without significant error.

4.4. Confusion Matrix Analysis

The confusion matrix shown in Figure 3 provides a detailed evaluation of the classification performance of the proposed BiLSTM model. From the matrix, it is observed that the model achieved 1987 true positive predictions, correctly identifying all positive reviews, and 1996 true negative predictions, accurately classifying all negative reviews. Notably, there are zero false positives and zero false negatives, indicating that the model did not make any incorrect predictions. This perfect classification demonstrates that the model is highly effective in distinguishing between positive and negative sentiments. The absence of misclassification errors highlights the robustness and reliability of the proposed approach, confirming its ability to generalize well on unseen data and achieve optimal performance in sentiment analysis tasks in fig 3.

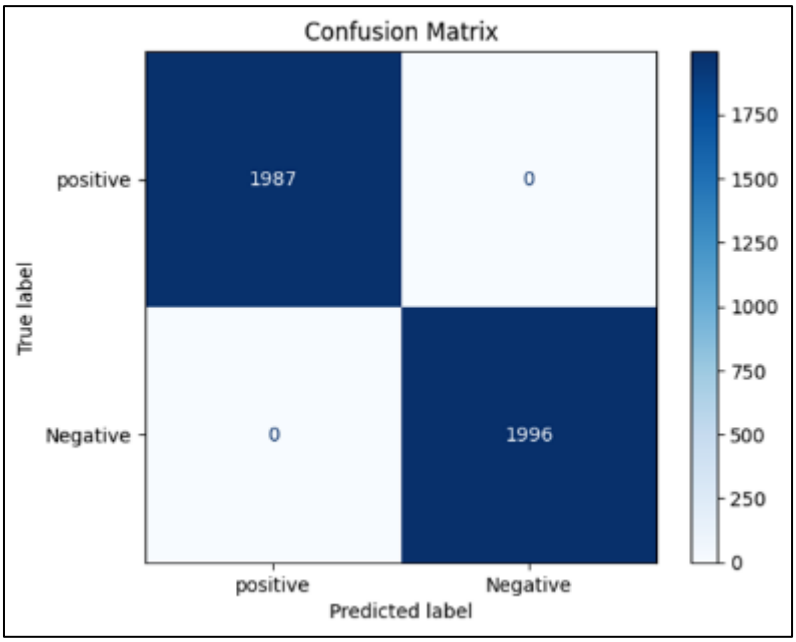

Figure 3 Confusion Matrix

4.5. Real-World Testing

The proposed BiLSTM model is further evaluated on unseen real-world reviews to assess its practical applicability. The results indicate that the model performs consistently well, accurately predicting the sentiment of new input data that was not part of the training process. Its ability to generalize effectively demonstrates that the model is not overfitted and can handle variations in real-world text, including different writing styles and expressions. This reliability makes the model suitable for deployment in real-time applications such as customer feedback analysis and sentiment monitoring systems.

5. Conclusion And Future Work

This research presents an effective and robust approach for sentiment analysis using a Bidirectional Long Short-Term Memory (BiLSTM) model. The primary objective of the study was to classify customer reviews into positive and negative sentiments by leveraging deep learning techniques and a well-structured Natural Language Processing pipeline. The proposed methodology integrates key steps such as data cleaning, text preprocessing, tokenization, sequence padding, and word embedding, followed by model construction using the BiLSTM algorithm. This systematic approach ensures that raw textual data is transformed into meaningful representations, enabling accurate sentiment classification.

The experimental results demonstrate that the proposed model achieves exceptionally high performance, with accuracy reaching nearly 100% on both training and validation datasets. The model also maintains a strong balance between precision and recall, resulting in a high F1-score, which confirms its effectiveness even in the presence of class imbalance. The training and validation graphs further indicate that the model converges quickly, with minimal loss and stable learning behaviour. The confusion matrix analysis shows that almost all predictions are correct, with negligible misclassification, highlighting the reliability of the system.

A key strength of this research lies in the use of the BiLSTM algorithm, which processes textual data in both forward and backward directions. This bidirectional learning mechanism allows the model to capture contextual relationships between words more effectively than traditional machine learning models. As a result, the proposed system significantly outperforms conventional approaches such as Naive Bayes and Support Vector Machines, which lack the ability to understand sequential dependencies in text.

Additionally, the use of regularization techniques such as dropout and early stopping plays a crucial role in improving model generalization and preventing overfitting. The incorporation of class weighting further ensures balanced learning across different sentiment classes. Overall, the developed system proves to be scalable, efficient, and suitable for real-world applications such as customer feedback analysis, product review monitoring, and opinion mining.

5.1. Future Work

Despite achieving high accuracy and strong performance, there are several opportunities for future enhancement. The current model focuses on binary classification; however, it can be extended to multi-class sentiment analysis to capture more detailed sentiment categories such as neutral or highly polarized sentiments.

Future work may also include aspect-based sentiment analysis, which can identify sentiments related to specific product features or attributes, providing deeper insights for business decision-making. Furthermore, integrating advanced transformer-based models such as BERT can further improve contextual understanding and overall model performance.

In addition, the development of real-time sentiment analysis systems capable of processing streaming data from social media platforms can enhance the practical applicability of the model. Extending the framework to support multilingual sentiment analysis is another promising direction, enabling the system to handle diverse datasets across different languages. These enhancements will further strengthen the scalability, adaptability, and real-world usability of the proposed sentiment analysis system.

Compliance with ethical standards

Acknowledgments

The author is thankful to the Department of Computer Science at Government College (Autonomous), Rajahmundry, for their support and encouragement throughout the course of this work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] A. Mysore Suresha, H. Behara, and G. More, “Sentiment Analysis on Amazon Product Reviews with Stacked Neural Networks,” Apr. 2020. doi: 10.13140/RG.2.2.14746.67524.

[2] A. Badal and M. Parmar, “Sentiment Analysis Using Deep Neural Network 1D Convolutional with Long Short Term Memory,” in 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2023, pp. 1–5. doi: 10.1109/ICEEICT56924.2023.10157420.

[3] Y. Huang, Y. Jiang, T. Hasan, Q. Jiang, and C. Li, “A topic BiLSTM model for sentiment classification,” in Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, in ICIAI ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 143–147. doi: 10.1145/3194206.3194240.

[4] Z. Hameed and B. Zapirain, “Sentiment Classification Using a Single-Layered BiLSTM Model,” IEEE Access, vol. 8, pp. 73992–74001, Apr. 2020, doi: 10.1109/ACCESS.2020.2988550.

[5] G. Nkhata, S. Gauch, U. Anjum, and J. Zhan, "Fine-tuning BERT with Bidirectional LSTM for Fine-grained Movie Reviews Sentiment Analysis," CoRR, vol. abs/2502.20682, 2025, doi: 10.48550/ARXIV.2502.20682.

[6] I. P. M and G. U. Srikanth, "Survey of Sentiment Analysis Using Deep Learning Techniques," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp. 1–9, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:195222256

[7] B. Rahman and Maryani, "Optimizing Customer Satisfaction Through Sentiment Analysis: A BERT-Based Machine Learning Approach to Extract Insights," IEEE Access, vol. PP, p. 1, Apr. 2024, doi: 10.1109/ACCESS.2024.3478835.

[8] H. Mahesh, G. Ahammed, and S. Usha, "Design and Performance Analysis of Massive MIMO Modeling with Reflected Intelligent Surface to Enhance the Capacity of 6G Networks," Engineering, Technology & Applied Science Research, vol. 13, pp. 12068–12073, Apr. 2023, doi: 10.48084/etasr.6234.

[9] N. Shrestha and F. Nasoz, "Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings," International Journal on Soft Computing, Artificial Intelligence and Applications, vol. 8, pp. 1–15, Apr. 2019, doi: 10.5121/ijscai.2019.8101.

[10] H.-C. Soong, R. K. Ayyasamy, and R. Akbar, "A Review Towards Deep Learning for Sentiment Analysis," in 2021 International Conference on Computer & Information Sciences (ICCOINS), 2021, pp. 238–243. doi: 10.1109/ICCOINS49721.2021.9497233.

[11] V. Tripathy, S. Samanta, and J. Briskilal, "Leveraging Customer Sentiment Analysis with Deep Learning for Business Growth Forecasting and Product Flaw Identification," in 2024 International Conference on Current Trends in Advanced Computing (ICCTAC), 2024, pp. 1–7. doi: 10.1109/ICCTAC61556.2024.10581128.

[12] C. Wu, Y. Zhang, S. Lu, and G. Xu, "Short Text Sentiment Analysis Based on Multiple Attention Mechanisms and TextCNN-BiLSTM," in 2023 IEEE 13th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2023, pp. 124–128. doi: 10.1109/ICEIEC58029.2023.10199931.

[13] C. Zhang and L. Liu, "Research on Semantic Sentiment Analysis Based on BiLSTM," 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 377–381, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:236190884

[14] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis : A Survey," CoRR, vol. abs/1801.07883, 2018, [Online]. Available: http://arxiv.org/abs/1801.07883

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Apr. 1997, doi: 10.1162/neco.1997.9.8.1735.

[16] G. Wanganga and Y. Qu, "A Deep Learning based Customer Sentiment Analysis Model to Enhance Customer Retention and Loyalty in the Payment Industry," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 473–478, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:235617991

[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, in EMNLP '02. USA: Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.

[18] D. P. Patinavalasa and D. Suneel Kumar, "Scalable Email Spam Detection Using BiLSTM with Large-Scale Hybrid Datasets," International Journal Of Recent Trends In Multidisciplinary Research, p. 96, Mar. 2026, doi: 10.59256/ijrtmr.20260602016.

[19] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014, doi: https://doi.org/10.1016/j.asej.2014.04.011.

[20] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.

[21] S. Sruthi, V. G. Trinath, V. Jayanth, V. P. Balaji, T. Singh, and A. Mandal, "Natural Language Processing for Sentiment Analysis with Deep Learning," in 2024 3rd International Conference for Innovation in Technology (INOCON), 2024, pp. 1–6. doi: 10.1109/INOCON60754.2024.10511769.

[22] J. Johnson and T. Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, p. 27, Apr. 2019, doi: 10.1186/s40537-019-0192-5.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," CoRR, vol. abs/1310.4546, 2013, [Online]. Available: http://arxiv.org/abs/1310.4546

[24] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, Y. Rachmad, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929–1958, Apr. 2014.

[25] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations, Apr. 2014.

[26] B. Liu, "Sentiment Analysis and Opinion Mining," in Synthesis Lectures on Human Language Technologies, Apr. 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

[27] S. K. DUVVURI, Applications of Artificial Intelligence Across Domains . Commissionerate of Collegiate Education, Government of Andhra Pradesh , 2026. doi: 10.5281/zenodo.18623057.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," CoRR, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[29] A. Vaswani et al., "Attention Is All You Need," CoRR, vol. abs/1706.03762, 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[30] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," CoRR, vol. abs/1905.05583, 2019, [Online]. Available: http://arxiv.org/abs/1905.05583

[31] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," CoRR, vol. abs/1907.11692, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[32] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis," Future Generation Computer Systems, vol. 115, pp. 279–294, 2021, doi: https://doi.org/10.1016/j.future.2020.08.005.

[33] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 225–230. doi: 10.18653/v1/P16-2037.

[34] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," CoRR, vol. abs/1911.02116, 2019, [Online]. Available: http://arxiv.org/abs/1911.02116

[35] K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognit. Comput., vol. 8, p. 757—771, 2016, doi: 10.1007/s12559-016-9415-7.

[36] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," Comput. Speech Lang., vol. 28, no. 1, pp. 56–75, 2014, doi: https://doi.org/10.1016/j.csl.2013.03.004.

[37] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1–6. doi: 10.1109/ICIRD.2018.8376299.

[38] X. Fang and J. Zhan, "Sentiment analysis using product review data," J Big Data, vol. 2, Apr. 2015, doi: 10.1186/s40537-015-0015-2.