International Journal of Scientific Research in Engineering & Technology
Volume6, Issue2 (March-April 2026), PP: 266-276.
https://www.doi.org/10.59256/ijsreat.20260602034
www.ijsreat.com
ISSN No: 2583-1240

Deep Learning-Based Sentiment Classification of Ride- Hailing Customer Reviews Using BiLSTM

Kanem Bharath Varma[1], Suneel Kumar Duvvuri[2]
[1]Student, M.Sc, Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.
[2]Assistant Professor, Department of Computer Science, Government College (Autonomous),, Rajahmundry, Andhra Pradesh, India.

To Cite this Article: *Kanem Bharath Varma[1], Suneel Kumar Duvvuri[2] , "Deep Learning-Based Sentiment Classification of Ride- Hailing Customer Reviews Using BiLSTM", International Journal of Scientific Research in Engineering & Technology, Volume 06, Issue 02, March-April 2026, PP: 266-276.*

Abstract: *Ride-hailing platforms such as Ola, Uber, Rapido, Lyft, Bolt, Grab, and other app-based transport services have become an important part of modern urban mobility. Customer reviews posted on these platforms contain useful opinions about service quality, driver behavior, booking experience, pricing, cancellation issues, payment problems, and overall user satisfaction. Since manual analysis of large review data is difficult, automated sentiment analysis is useful for understanding customer feedback. This project presents a deep learning-based binary sentiment analysis framework for classifying ride-hailing customer reviews into positive and negative categories. The original dataset contained 12,000 reviews with review text and rating scores. After removing neutral ratings, the final dataset contained 11,667 reviews, including 8,732 positive reviews and 2,935 negative reviews. Reviews with ratings 4 and 5 were labelled as positive, while ratings 1 and 2 were labelled as negative. The review text was cleaned by applying lowercase conversion, URL removal, special character removal, stopword removal, tokenization, and sequence padding. The proposed model was developed using a Bidirectional Long Short-Term Memory network. The model includes an embedding layer, SpatialDropout1D layer, BiLSTM layer, dropout layer, dense layer, and sigmoid output layer for binary classification. The dataset was divided into 80% training data and 20% testing data. The model was trained using binary cross-entropy loss and Adam optimizer with early stopping and learning-rate reduction.*

Key Words: *Ride-hailing platforms, customer reviews, sentiment analysis, binary classification, deep learning, BiLSTM, text Preprocessing, customer satisfaction.*

I. INTRODUCTION

Ride-hailing services have become an important part of modern transportation systems because they provide convenient, flexible, and fast access to mobility through mobile applications. Platforms such as Uber, Ola, Rapido, Lyft, Bolt, Grab, and other app-based transport services have changed the way people travel by offering features such as online ride booking, real-time driver tracking, fare visibility, route information, digital payment, and post-ride feedback. These services are widely used in urban areas because they reduce the difficulty of finding transportation and provide users with quick access to travel services through smartphones [1], [2].

As ride-hailing platforms continue to grow, customer feedback has become an important source of service-related information. After completing a ride, users often provide ratings and written reviews based on their experience. These reviews may include opinions about driver behavior, waiting time, booking process, fare charges, cancellation issues, payment experience, app performance, safety, comfort, and customer support. Such reviews are valuable because they represent real user experiences and can help identify whether customers are satisfied or dissatisfied with the service [3], [4]. Customer reviews are more informative than rating scores alone. A numerical rating may show whether a customer had a positive or negative experience, but the written review explains the reason behind that rating. For example, a customer may give a low rating because of driver cancellation, high fare, delayed pickup, or poor customer support. Similarly, a positive review may reflect good driver behavior, smooth booking, affordable pricing, or comfortable ride experience. Therefore, review text provides deeper insight into customer perception and service quality [5][6]

However, manually reading and analyzing a large number of customer reviews is difficult. Ride-hailing applications receive thousands of reviews across platforms, and these reviews are written in different styles, lengths, and formats. Some reviews are short and direct, while others are longer and more descriptive. Many reviews may contain informal language, spelling variations, abbreviations, punctuation errors, or incomplete sentences. Because of these challenges, manual sentiment classification is time-consuming and may not produce consistent result [7] [8]

Sentiment analysis provides an automated solution to this problem. Sentiment analysis is a Natural Language Processing technique used to identify opinions, emotions, or attitudes expressed in text. In customer review analysis, it is commonly used to classify reviews into positive, negative, or neutral categories. For this study, binary sentiment classification is used, where ride-

hailing customer reviews are classified as either positive or negative. This approach helps convert large-scale unstructured customer feedback into meaningful sentiment categories [9], [10].

Traditional machine learning methods such as Naive Bayes, Support Vector Machine, and Random Forest have been widely used for sentiment classification [11], [12]. Although these methods can provide useful results, they often depend on manually selected features and may not fully capture word order or contextual meaning [13]. Customer reviews are sequential in nature, where the meaning of a word may depend on surrounding words. For example, the phrase "not good" has a negative meaning, even though the word "good" alone is positive. Therefore, context-aware deep learning models are useful for sentiment analysis [14].

Deep learning models can automatically learn useful patterns from textual data [15]. Among them, Long Short-Term Memory networks are suitable for sequence-based text classification because they can capture dependencies across time steps [16]. Bidirectional Long Short-Term Memory improves this process by reading the input sequence in both forward and backward directions [17]. This allows the model to understand both previous and future word context in a review. Hence, BiLSTM is suitable for ride-hailing customer review sentiment classification [18], [19].

II.LITERATURE REVIEW

This prospective comparative study was carried out on patients of Department of general Medicine at Dr. Ram Manohar Lohia The literature review provides an understanding of the major studies already conducted in the field of ride-hailing review sentiment analysis. It helps in identifying the methods used by previous researchers, the performance achieved by different models, and the gaps that still remain in this area. Through this review, the present study is positioned within the existing body of research and its relevance is clearly established.

Sentiment analysis of ride-hailing and transportation application reviews has become an important research area because customer reviews provide direct information about service quality, user satisfaction, and operational issues. In app-based transportation platforms, users frequently express their experiences through ratings and written reviews after using the service. These reviews reflect practical aspects such as booking convenience, driver behavior, waiting time, fare charges, safety, ride comfort, and app performance. Therefore, customer review sentiment analysis helps in understanding how users perceive ride-hailing services and in identifying major strengths and weaknesses of such platforms.

Several previous studies have applied machine learning methods to ride-hailing and transportation application reviews. Hermanto et al. [20] analyzed Gojek and Grab user reviews using Support Vector Machine with Particle Swarm Optimization and reported 73.09% accuracy. Kurniawati et al. [21] also applied Support Vector Machine to Maxim application reviews and achieved 89.82% accuracy. These studies show that traditional supervised machine learning methods can be effectively used for transportation application review classification.

Other studies adopted probabilistic and comparative machine learning approaches. Rahman et al. [22] studied Gojek application reviews using Naive Bayes and reported 76% accuracy, along with weighted precision of 82% and F1-score of 78%. In another study, Ahammad et al. [23] conducted a comparative analysis of ride-sharing application reviews using both machine learning and deep learning models, where Support Vector Machine achieved 76.70% accuracy. These findings indicate that machine learning models can produce useful classification results, although their performance often depends on the quality of preprocessing, the nature of the dataset, and the feature representation method used.

Some studies focused on improving classification performance through app-specific analysis and model comparison. Saefullah et al. [24] analyzed Maxim app user reviews in Indonesia using multiple machine learning models and reported that Support Vector Machine achieved 91.3% accuracy, while Random Forest achieved 89% and Naive Bayes achieved 78%. The study used 5,000 Google Play reviews, applied preprocessing and TF-IDF features, and used SMOTE to address class imbalance, showing that both algorithm choice and data preparation strategy can strongly influence sentiment classification performance. Likewise, Romadhoni et al. [25] analyzed InDrive review data and compared Support Vector Machine and Naive Bayes, where Support Vector Machine achieved 78% accuracy and Naive Bayes achieved 76% accuracy. These studies show that both model selection and data preparation strategy can significantly influence sentiment classification performance in ride-hailing application reviews.

More recent studies have moved toward stronger comparative and advanced approaches. Kristiyanto et al. [26] reported 91% accuracy with Random Forest and 89% accuracy with Support Vector Machine in a study based on Gojek reviews. In another advanced study, Safitri et al. [27] applied BERT-Base Multilingual Uncased to Maxim application reviews and achieved 94.7% accuracy. These results show that advanced and contextual models can improve the effectiveness of ride- hailing review sentiment analysis and may outperform many traditional machine learning methods.

Although previous studies achieved useful results, there is still a need for a focused BiLSTM-based binary sentiment classification framework for ride-hailing customer reviews. Many earlier works used traditional machine learning models that may not fully capture the sequential meaning of review text. Some advanced transformer-based models achieved strong performance, but they may require greater computational resources and more complex implementation settings. BiLSTM provides a balanced approach because it captures contextual information from both forward and backward directions while remaining suitable for structured deep learning-based review classification. The comparative summary of these studies is presented in Table 1.

Author(s)	Year	Platform / Data Focus	Method / Model Used	Reported Accuracy / Performance
Hermanto et al. [20]	2020	Gojek and Grab user reviews	SVM with Particle Swarm Optimization	73.09% accuracy

Kurniawati et al. [21]	2023	Maxim app reviews	Support Vector Machine	89.82% accuracy
Rahman et al. [22]	2024	Gojek app reviews	Naive Bayes	76% accuracy, weighted precision 82%, F1-score 78%
Ahammad et al. [23]	2024	Various ride-sharing app reviews	Comparative machine learning and deep learning models	SVM: 76.70% accuracy
Saefullah et al. [24]	2024	Maxim app user reviews in Indonesia	SVM, Random Forest, and Naive Bayes	SVM: 91.3%, Random Forest: 89%, Naive Bayes: 78%
Romadhoni et al. [25]	2025	InDrive app reviews	Support Vector Machine and Naive Bayes	SVM: 78% accuracy; Naive Bayes: 76% accuracy
Kristiyanto et al. [26]	2025	Gojek user reviews	Random Forest and SVM	Random Forest: 91% accuracy; SVM: 89% accuracy
Safitri et al. [27]	2025	Maxim app reviews	BERT-Base Multilingual Uncased	94.7% accuracy
Proposed Study	2026	Uber customer reviews	BiLSTM	94.73% accuracy

Table 1. Comparative Analysis of Previous Studies on Ride-Hailing Review Sentiment Analysis

III.METHODOLOGY

3.1 System Overview

The proposed system performs binary sentiment classification of ride-hailing customer reviews using a BiLSTM model. The main aim is to classify each review as either positive or negative based on the customer's written opinion. The system follows a Natural Language Processing workflow that includes dataset loading, data cleaning, label creation, text preprocessing, tokenization, sequence padding, model training, and evaluation.

The overall workflow adopted in this study is consistent with standard sentiment analysis and opinion mining procedures used for transforming opinion-oriented text into sentiment categories [28][[29]

In this study, the content column is used as the review text, and the score column is used to create sentiment labels. Reviews with scores 1 and 2 are considered negative, while reviews with scores 4 and 5 are considered positive. Reviews with score 3 are removed to maintain a clear binary classification structure. After filtering, the final dataset contains 11,667 reviews, including 8,732 positive reviews and 2,935 negative reviews.

3.2 Research Design

This study follows an experimental research design. The dataset is processed through a systematic computational procedure, and the BiLSTM model is trained and tested on the prepared review data. The design is suitable because sentiment classification requires a clear step-by-step process from raw review text to final sentiment prediction.

The process begins with dataset inspection and quality checking. After that, rating scores are converted into binary labels. The review text is cleaned and transformed into numerical sequences using tokenization and padding. Finally, the BiLSTM model is trained and evaluated using standard classification metrics.

3.3 Development Environment and Tools Used

The implementation is carried out using Python in Google Colab or Jupyter Notebook. Python libraries such as Pandas and NumPy are used for data handling, Regular Expressions and NLTK are used for text preprocessing, and Matplotlib, Seaborn, and WordCloud are used for visualization. Scikit-learn is used for train-test splitting and evaluation, while TensorFlow/Keras is used for building and training the BiLSTM model in table 2.

Tool / Library	Purpose
Python	Main programming language
Pandas, NumPy	Dataset handling and numerical operations
NLTK, Regular Expressions	Text cleaning and stopword handling
Matplotlib, Seaborn, WordCloud	Visualization
Scikit-learn	Train-test split and evaluation
TensorFlow / Keras	Deep learning model development

Table 2. Development Tools and Purpose

3.4 Dataset Description

The dataset used in this study is a ride-hailing customer review dataset. The original dataset contains 12,000 records and 10 columns. However, only two columns are used for this study: content and score. The content column contains the written customer review, while the score column contains the rating value.

The review text is used as the input feature, and the score is used to generate the sentiment label. After removing score 3 reviews and invalid records, the final dataset contains 11,667 reviews. This dataset is suitable because it contains customer opinions related to ride-hailing service experience, including driver behavior, ride comfort, pricing, booking experience, and app performance in table 3.

Parameter	Description
Dataset Type	Ride-hailing customer review dataset
Original Dataset Size	12,000 reviews
Selected Columns	Content and score
Final Dataset Size	11,667 reviews
Input Feature	Customer review text
Target Feature	Sentiment label
Classification Type	Binary sentiment classification

Table 3. Dataset Description

3.5 Data Inspection and Quality Checking

Data inspection is performed before preprocessing to check the structure and quality of the dataset. The column names, data types, number of records, and sample reviews are examined. Since the study focuses only on sentiment classification, unnecessary columns are removed, and only the review text and rating score are retained.

Rows with missing values, empty review text, or invalid records are removed. This step is important because incomplete or noisy data may affect the performance of the model. After quality checking, the dataset becomes cleaner and more suitable for sentiment classification.

3.6 Class Distribution Analysis

Class distribution analysis is used to understand the number of positive and negative reviews. After removing score 3 reviews, scores 1 and 2 are mapped as negative, and scores 4 and 5 are mapped as positive. The final dataset contains more positive reviews than negative reviews in table 4.

Sentiment Class	Number of Reviews
Positive	8,732
Negative	2,935
Total	11,667

Table 4. Final Class Distribution

This distribution shows that the dataset is imbalanced, with more positive reviews. To handle this during splitting, stratified train-test splitting is used so that both training and testing sets maintain a similar class ratio.

3.7 Label Encoding

Label encoding is used to convert sentiment labels into numerical values. Since deep learning models cannot directly process text labels such as positive and negative, the labels are converted into binary form. In this study, negative sentiment is encoded as 0, and positive sentiment is encoded as 1. This format is suitable for binary classification using sigmoid activation in the output layer.

3.8 Text Cleaning and Preprocessing

Text preprocessing is applied to clean the raw customer reviews. Reviews may contain URLs, special characters, punctuation marks, numbers, emojis, uppercase letters, and extra spaces. These elements may reduce the quality of model learning if they are not handled properly.

Text preprocessing is an essential step in sentiment analysis because online review text often contains noisy and unstructured elements that must be standardized before classification [30].

In this study, the text is converted into lowercase, URLs are removed, special characters and punctuation marks are removed, and stopwords are filtered. However, important negation words such as “not,” “no,” “nor,” and “but” are retained because they can change the sentiment meaning of a review. After cleaning, the processed text is stored as clean_review. in Fig 1.

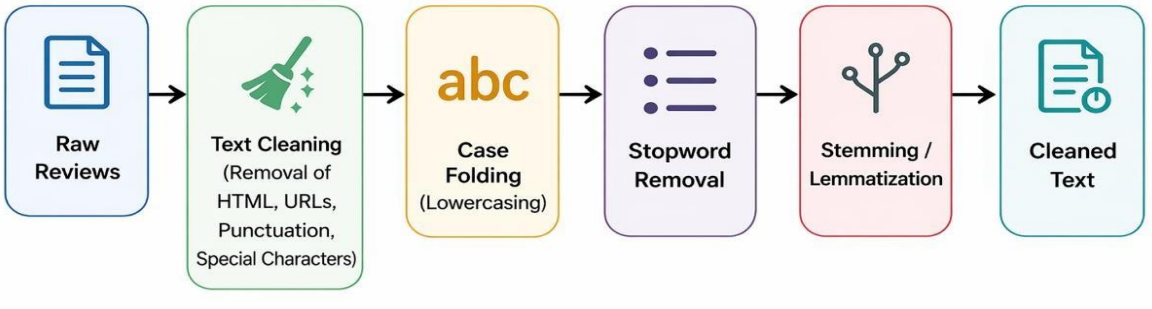

Fig 1. Data Preprocessing Pipeline

3.9 Feature Representation

Feature representation converts the cleaned review text into numerical form. Since the BiLSTM model cannot directly process raw text, tokenization and sequence padding are applied. Tokenization converts words into integer values, and padding makes all review sequences equal in length.

3.9.1 Tokenization

Tokenization assigns a numerical index to each word in the cleaned review text. In this study, the vocabulary size is fixed at 10,000 words. The tokenizer is fitted on the training data and then used to convert both training and testing reviews into integer sequences.

3.9.2 Sequence Padding

After tokenization, reviews may have different sequence lengths. Therefore, padding is applied to make all review sequences equal in length. In this study, the maximum sequence length is fixed at 100. Short reviews are padded with zeros, while long reviews are truncated. The final padded training data shape is (9333, 100), and the testing data shape is (2334, 100).

3.10 Train-Test Split

The dataset is divided into training and testing sets using an 80:20 ratio. The training set contains 9,333 reviews, and the testing set contains 2,334 reviews. Stratified splitting is used to maintain a similar positive and negative class distribution in both sets. The training data is used to train the model, while the testing data is used to evaluate the final performance.

3.11 LSTM Cell Structure Used in BiLSTM

A separate standalone LSTM model is not implemented in this study. However, the proposed BiLSTM model is built using LSTM units. Therefore, the LSTM cell structure is briefly explained to understand how the BiLSTM processes review sequences.

Sequence models in natural language processing depend on learned numerical word representations, which later support contextual sequence learning in recurrent architectures [31].

An LSTM cell uses gates to control the flow of information. These gates help the model decide what information should be remembered, updated, or removed while processing the review text.

3.11.1 Embedding Layer

The embedding layer converts tokenized word indices into dense vector representations. In this study, the vocabulary size is 10,000, the embedding dimension is 128, and the input length is 100. This layer helps the model learn meaningful word representations before passing them to the BiLSTM layer.

3.11.2 LSTM Cell Equations

The LSTM cell uses different gates to process sequence information.

The gated memory mechanism used in these equations is derived from the original Long Short-Term Memory architecture, which was developed to preserve sequence information more effectively across time steps [32].

Forget Gate

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

The forget gate decides how much previous information should be retained or removed.

Input Gate

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

The input gate decides how much new information should be added to the memory cell.

Candidate State

$$\tilde{c}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (3)$$

The candidate state creates new possible information for the cell state.

Cell State Update

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

The cell state update combines retained old information with new candidate information.

Output Gate

$$o_t = \sigma(W_O[h_{t-1}, x_t] + b_O) \quad (5)$$

The output gate controls how much information is passed to the hidden state.

Hidden State

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

The hidden state represents the output of the LSTM cell at time step *t*.

3.12 Proposed BiLSTM Model

The proposed model is based on Bidirectional Long Short-Term Memory. BiLSTM processes the input review sequence in both forward and backward directions. This is useful because the sentiment meaning of a word may depend on both previous and future words in the review.

In the forward direction, the model reads the review from the first word to the last word. In the backward direction, the model reads the same review from the last word to the first word. The hidden states from both directions are combined to create a stronger contextual representation in Fig2.

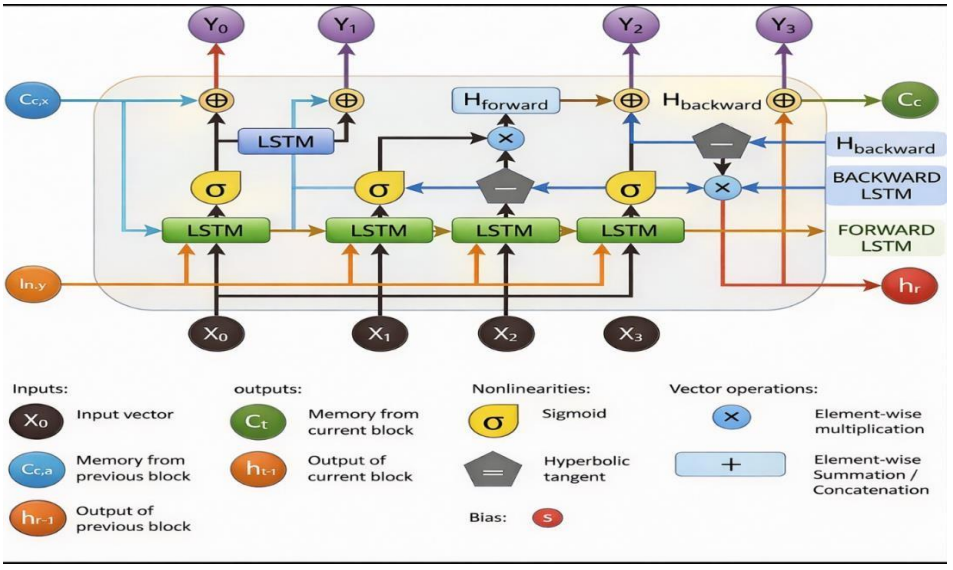

Fig 2. Architecture of the Proposed BiLSTM Model

3.12.1 Forward Pass

$$\overrightarrow{h_t} = \mathrm{LSTM}_f(x_t, \overrightarrow{h_{t-1}}) \quad (7)$$

The forward pass processes the review sequence from left to right and captures information from previous words.

3.12.2 Backward Pass

$$\overleftarrow{h_t} = \mathrm{LSTM}_b(x_t, \overleftarrow{h_{t+1}}) \quad (8)$$

The backward pass processes the review sequence from right to left and captures information from future words.

3.12.3 Final BiLSTM Output

$$y_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (9)$$

The final BiLSTM output is obtained by combining the forward and backward hidden states. This combined representation is passed to the dense layers for final sentiment classification.

This bidirectional combination provides a richer contextual sequence representation for sentiment classification [33].

3.13 Model Configuration

The proposed BiLSTM model includes an embedding layer, SpatialDropout1D layer, Bidirectional LSTM layer, dropout layer, dense layer, and sigmoid output layer in Table 5.

Layer	Configuration	Purpose
Embedding Layer	Vocabulary size = 10,000; Dimension = 128	Converts word indices into dense vectors
SpatialDropout1D	Dropout rate = 0.3	Reduces overfitting
Bidirectional LSTM	64 units	Captures forward and backward context
Dropout Layer	Dropout rate = 0.4	Improves generalization
Dense Layer	32 neurons; ReLU activation	Learns sentiment features
Output Layer	1 neuron; Sigmoid activation	Produces binary sentiment output

Table 5. Proposed BiLSTM Model Configuration

3.14 Training Strategy

The BiLSTM model is trained using tokenized and padded review sequences. The model is trained for up to 10 epochs with a batch size of 64. A validation split is used during training to monitor performance. Early stopping is applied to stop training when validation loss does not improve, and ReduceLROnPlateau is used to reduce the learning rate when improvement slows down. These techniques help reduce overfitting and improve training stability.

3.15 Loss Function and Optimizer

Binary cross-entropy is used as the loss function because the task is binary sentiment classification.

$$\text{Loss} = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \quad (10)$$

Adam optimizer is used to train the model with a learning rate of 0.001. The model is compiled using binary cross-entropy loss, Adam optimizer, and accuracy as the training metric.

3.16 Evaluation Metrics

The model performance was evaluated using accuracy, precision, recall, and F1-score. These metrics provide a balanced understanding of spam classification results.

Accuracy: Accuracy measures the overall proportion of correctly classified messages out of the total number of messages.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Precision: Precision measures the proportion of correctly predicted spam messages among all messages predicted as spam.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

Recall: Recall measures the proportion of actual spam messages that are correctly identified by the model

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

F1-score: F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification Performance

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \quad (14)$$

Pseudo algorithm

The complete workflow of the proposed ride-hailing customer review sentiment classification framework is summarized in Algorithm.

Algorithm: BiLSTM-Based Sentiment Classification of Ride-Hailing Customer Reviews Input: Ride-hailing customer review dataset

Output: Sentiment class as positive or negative

1. Load the ride-hailing customer review dataset.
2. Select the required columns: content and score.
3. Remove missing, empty, and invalid review records.
4. Remove neutral reviews with score 3.
5. Convert rating scores into sentiment labels:

- Scores 1 and 2 → Negative
- Scores 4 and 5 → Positive

6. Encode sentiment labels:

- Negative → 0
- Positive → 1

7. Clean the review text using lowercase conversion, URL removal, special character removal, punctuation removal, and stopword removal.
8. Split the dataset into training and testing sets using an 80:20 ratio.
9. Apply tokenization and sequence padding to convert text into numerical input.
10. Build and train the BiLSTM model using embedding, BiLSTM, dropout, dense, and sigmoid output layers.
11. Evaluate the model using testing data.
12. Calculate accuracy, precision, recall, F1-score, test loss, and confusion matrix.
13. Display the final sentiment classification result.

3.17 Summary of Methodology

This methodology explains the complete process used for binary sentiment classification of ride-hailing customer reviews. The dataset is first loaded and inspected. The required columns are selected, invalid records are removed, and rating scores are converted into positive and negative sentiment labels. After preprocessing, the review text is converted into numerical sequences using tokenization and padding.

The proposed BiLSTM model processes the review sequences in both forward and backward directions to capture contextual meaning. The model is trained using binary cross-entropy loss and Adam optimizer. Finally, the model is evaluated using accuracy, precision, recall, F1-score, test loss, and confusion matrix. This methodology provides a clear framework foranalyzing ride-hailing customer reviews and predicting customer sentiment.

IV.RESULT AND DISCUSSION

4.1 Experimental Setup

The proposed ride-hailing customer review sentiment classification model was implemented using Python with deep learning libraries. The dataset was divided into training and testing sets using an 80:20 ratio. The proposed BiLSTM model was trained on the prepared review dataset to classify customer reviews into positive and negative sentiment categories. The main training parameters included 10 epochs, batch size of 64, validation split of 0.1, Adam optimizer, binary cross- entropy loss function, early stopping, and learning-rate reduction. This setup helped the model learn efficiently while reducing the possibility of overfitting during training.

The original dataset contained 12,000 ride-hailing customer reviews. After removing neutral reviews with score 3 and preparing the dataset for binary classification, the final dataset contained 11,667 reviews. Among these, 8,732 reviews were positive and 2,935 reviews were negative. The cleaned reviews were converted into numerical sequences using tokenization with a vocabulary size of 10,000 and padding to a fixed sequence length of 100. This transformed the raw review text into a structured format suitable for BiLSTM-based sentiment classification.

4.2 Preliminary Data Analysis

Before model training, the dataset was carefully inspected and pre-processed to ensure its suitability for sentiment classification. Only the required columns, namely review content and rating score, were selected for the study. Reviews with missing text, invalid records, and neutral score values were removed. The rating scores were then converted into binary sentiment labels, where scores 1 and 2 were treated as negative and scores 4 and 5 were treated as positive.

After preprocessing, the final dataset contained 11,667 customer reviews. The class distribution showed that positive reviews were higher than negative reviews. This indicates that many users expressed satisfaction with the ride-hailing service. However, the negative reviews were also important because they contained complaints about driver behavior, ride cancellation, waiting time, fare charges, payment problems, and poor service experience.

Review length analysis was also performed to understand the nature of the text data. The mean number of characters was about 62.60, the mean number of words was about 11.62, and the mean number of sentences was about 1.50. These values show that most ride-hailing customer reviews were short and direct. Therefore, proper preprocessing, tokenization, and sequence-based learning were important for effective sentiment classification.

4.3 Training Performance

The BiLSTM model showed strong learning behavior during training. In the first epoch, the model achieved a training accuracy of 0.8734 and a validation accuracy of 0.9465. In the second epoch, the training accuracy improved to 0.9558, while the validation accuracy reached 0.9572. This improvement indicates that the model learned useful sentiment patterns from the customer review text.

The loss values also reduced during training, showing that the model was able to minimize prediction error. The model achieved a training loss of 0.3023 in the first epoch and reduced it to 0.1306 in the second epoch. Early stopping and learning-rate reduction helped stabilize training when validation loss stopped improving. These observations indicate that the BiLSTM model learned the review patterns effectively and converged in a stable manner without major training instability.

The training behavior of the proposed model can be illustrated through the training and validation accuracy and loss curves shown in Figure 3.

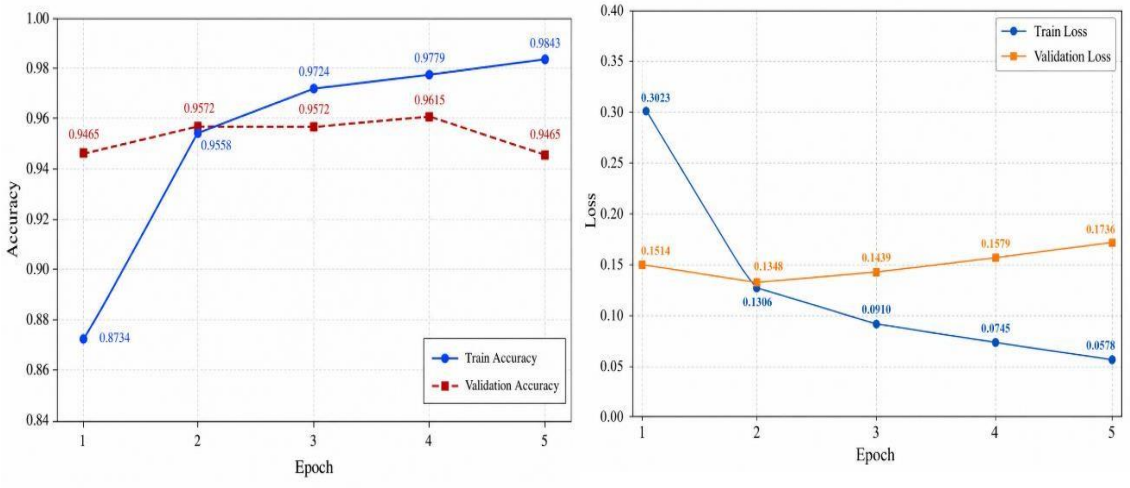

Fig 3. BiLSTM Model Performance Curves Showing Accuracy and Loss During Training

4.5 Confusion Matrix Analysis

The confusion matrix provides a detailed view of the classification performance of the proposed BiLSTM model on the test dataset. As shown in Figure 4.2, the model correctly classified a large number of both negative and positive reviews, asindicated by the higher values along the main diagonal of the confusion matrix in Fig 4.

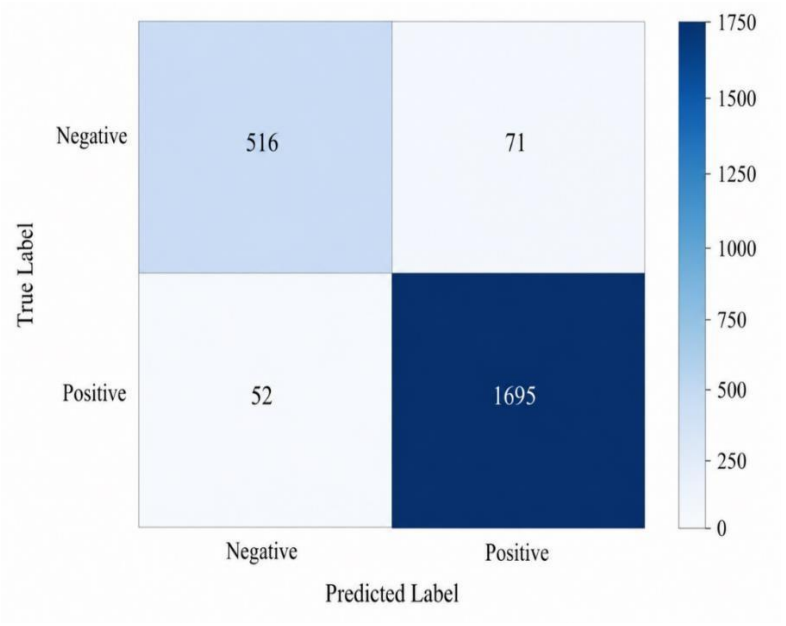

Fig 4. Confusion Matrix of the BiLSTM Model on the Test Dataset

From the confusion matrix, 516 negative reviews were correctly classified as negative, while 71 negative reviews were incorrectly classified as positive. Similarly, 1695 positive reviews were correctly classified as positive, while 52 positive reviews were incorrectly classified as negative. These results show that the model performed strongly in identifying both sentiment classes.

The test dataset contained 587 negative reviews and 1747 positive reviews. For the negative class, the model achieved a precision of 0.9085, recall of 0.8790, and F1-score of 0.8935. For the positive class, the model achieved a precision of 0.9598, recall of 0.9702, and F1-score of 0.9650. The overall accuracy obtained from the confusion matrix was 94.73%.

The confusion matrix shows that the proposed BiLSTM model performed slightly better on the positive class than on the negative class. This may be due to the larger number of positive reviews in the dataset. However, the model also showed good performance in identifying negative reviews, which is important because negative reviews often contain complaints related to fare charges, ride cancellations, waiting time, payment issues, driver behavior, and poor service experience.

Overall, the confusion matrix confirms that the proposed BiLSTM model is effective for binary sentiment classification of ride-hailing customer reviews.

4.6 Evaluation Metrics

In addition to confusion matrix analysis, the performance of the proposed BiLSTM model was evaluated using accuracy, precision, recall, and F1-score. These metrics provide a clear understanding of the model's overall classification ability on ride-hailing customer reviews.

The proposed BiLSTM model achieved an accuracy of 0.9473, precision of 0.9469, recall of 0.9473, and F1-score of 0.9470. These values show that the model performed strongly in classifying customer reviews into positive and negative sentiment categories. The accuracy value indicates that most of the test reviews were correctly classified. The precision value shows that the predicted sentiment labels were reliable. The recall value indicates that the model successfully identified the actual sentiment classes, while the F1-score shows a good balance between precision and recall in Table 6.

Metric	**BiLSTM**
Accuracy	0.9473
Precision	0.9469
Recall	0.9473
F1-Score	0.9470

Table 6. Evaluation Metrics of the BiLSTM Model

Overall, the evaluation metrics confirm that the proposed BiLSTM model provides effective and reliable performance for binary sentiment classification of ride-hailing customer reviews.

4.7 Discussion

The overall results show that the proposed BiLSTM model is effective for binary sentiment classification of ride-hailing customer reviews. The model achieved a model accuracy of 94.73%, with precision of 0.9469, recall of 0.9473, and F1- score

of 0.9470. These results indicate that the model can classify most customer reviews correctly into positive and negative sentiment categories.

The strong model accuracy shows that the BiLSTM model learned meaningful sentiment patterns from the review text. Since ride-hailing customer reviews are often short and context-dependent, the bidirectional learning ability of BiLSTM helped the model understand both previous and future word context.

The confusion matrix also supports the model performance. The model correctly classified 516 negative reviews and 1695 positive reviews. Although some misclassifications occurred, the number of correct predictions was much higher than the incorrect predictions. Overall, the achieved model accuracy of 94.73% confirms that the proposed BiLSTM framework is suitable for ride-hailing customer review sentiment

V.CONCLUSION AND FUTURE WORK

This research presented a deep learning-based sentiment classification framework for ride-hailing customer reviews using a Bidirectional Long Short-Term Memory model. The study focused on binary sentiment classification, where reviews were classified as positive or negative based on rating-derived sentiment labels.

The original dataset contained 12,000 customer reviews. After removing neutral reviews and invalid records, the final dataset contained 11,667 reviews, including 8,732 positive reviews and 2,935 negative reviews. The review text was cleaned through preprocessing steps such as lowercase conversion, URL removal, special character removal, punctuation removal, selective stopword handling, tokenization, and sequence padding.

The proposed BiLSTM model was built using an embedding layer, SpatialDropout1D layer, bidirectional LSTM layer, dropout layer, dense hidden layer, and sigmoid output layer. The model was trained using binary cross-entropy loss and Adam optimizer. The dataset was split into 80% training data and 20% testing data.

The experimental results showed that the proposed model achieved 94.73% test accuracy, 94.68% weighted precision, 94.73% recall, and 94.70% F1-score. These results confirm that the BiLSTM model can effectively learn sentiment patterns from ride-hailing customer review text. The study concludes that BiLSTM is a suitable and effective approach for binary sentiment classification in ride-hailing customer feedback analysis.

Future Scope

Although the proposed BiLSTM model achieved strong performance, the study can be extended in several ways. Future work may include multi-class sentiment classification by considering positive, negative, and neutral reviews. The study can also be expanded by using aspect-based sentiment analysis to identify specific service issues such as pricing, driver behavior, waiting time, safety, and app performance.

Future research may also compare BiLSTM with transformer-based models such as BERT and RoBERTa. In addition, multilingual ride-hailing reviews can be included to study customer opinions across different languages and regions. A larger dataset from multiple ride-hailing platforms may further improve the generalizability of the model.

REFERENCES

1. Z. Zulkarnain, I. Surjandari, and R. Wayasti, “Sentiment Analysis for Mining Customer Opinion on Twitter: A Case Study of Ride-Hailing Service Provider,” Apr. 2018, pp. 512–516. doi: 10.1109/ICISCE.2018.00113.
2. I. Surjandari, R. A. Wayasti, E. Laoh, Zulkarnain, A. M. M. Rus, and I. Prawiradinata, “Mining public opinion on ride-hailing service providers using aspect-based sentiment analysis,” *International Journal of Technology*, vol. 10, no. 4, pp. 818–828, Jul. 2019, doi: 10.14716/ijtech.v10i4.2860.
3. J. H. Kim, D. Nan, Y. Kim, and H. P. Min, “Computing the User Experience via Big Data Analysis: A Case of Uber Services,” *Computers, Materials and Continua*, vol. 67, no. 3, pp. 2819–2829, Mar. 2021, doi: 10.32604/cmc.2021.014922.
4. M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” 2004.
5. D.-H. Park, J. Lee, and I. Han, “The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement,” *International Journal of Electronic Commerce*, vol. 11, no. 4, pp. 125–148, 2007, doi: 10.2753/JEC1086-4415110405.
6. Pandiri Lavanya, Patinavalasa Durga Prasad, and Suneel Kumar Duvvuri, “Context-Aware Sentiment Classification of Movie Reviews Using Bidirectional LSTM Networks,” *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 13, no. 2, pp. 159–171, Mar. 2026, doi: 10.32628/IJSRSET261371.
7. S. K. DUVVURI, *Applications of Artificial Intelligence Across Domains* . Commissionerate of Collegiate Education, Government of Andhra Pradesh , 2026. doi: 10.5281/zenodo.18623057.
8. 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.
9. N. Malik and M. Bilal, “Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges,” *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2203.
10. F. Sebastiani, “Machine Learning in Automated Text Categorization,” 2002.
11. K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, “Text Classification Algorithms: A Survey,” 2019.
12. P. Amri, D. M. Suri, and Syuhada, “The Analysis of Ride Hailing User Characteristics from App Reviews,” 2024.
13. N. Fragkos and others, “A Sentiment Analysis Approach for Exploring Customer Experience and Service Quality Through Online Reviews,” 2024.
14. B. Pang, L. Lee, and S. Vaithyanathan, “Sentiment Classification Using Machine Learning Techniques,” 2002.
15. T. Nasukawa and J. Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing,” 2003.
16. B. Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” 2012.
17. R. Feldman, “Techniques and Applications for Sentiment Analysis,” 2013.
18. M. Alzate, M. Arce-Urriza, and J. Cebollada, “Mining the Text of Online Consumer Reviews to Analyze Brand Image and Brand

Positioning," 2022.
19. A. Karasenko and D. Baier, "Beyond Sentiment Analysis of Online Customer Reviews," 2025.
20. H. Hermanto and others, "Sentiment Analysis on Gojek and Grab User Reviews Using SVM Algorithm Based on Particle Swarm Optimization," 2020.
21. P. Kurniawati, R. Y. Fa'rifah, and D. Witarsyah, "Sentiment Analysis of Maxim Online Transportation App Reviews Using Support Vector Machine (SVM) Algorithm," 2023.
22. Z. Rahman and others, "Sentiment Analysis of Gojek App Reviews on Google Play Store with Natural Language Processing Using Naive Bayes Algorithm," 2024.
23. M. S. Ahammad and others, "Sentiment Analysis of Various Ride Sharing Applications Reviews: A Comparative Analysis Between Deep Learning and Machine Learning Algorithms," 2024.
24. R. Saefullah, S. Luthfi, O. Yohandoko, and A. Prabowo, "Sentiment Analysis of Maxim App User Reviews in Indonesia Using Machine Learning Model Performance Comparison," *International Journal of Quantitative Research and Modeling*, vol. 5, no. 3, pp. 331–340, 2024.
25. A. A. Romadhoni, A. Rachmadany, and B. H. Prasojo, "Sentiment Analysis of InDrive App Usage Reviews on Google Playstore Using Support Vector Machine (SVM) and Naïve Bayes Algorithm," 2025.
26. Kristiyanto and Sandiva, "Comparison of Random Forest and Support Vector Machine Learning Algorithms in Sentiment Analysis of Gojek User Reviews," 2026.
27. S. E. Safitri and others, "User Opinion Mining on the Maxim Application Reviews Using BERT-Base Multilingual Uncased," 2025.
28. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," 2008.
29. K. Sirisha, "Contextual Fake Review Detection in E-commerce using Bidirectional LSTM and Word Embeddings," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 14, no. 4, pp. 6336–6346, Apr. 2026, doi: 10.22214/ijraset.2026.80060.
30. B. Liu, *Sentiment Analysis and Opinion Mining*. 2012.
31. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.
32. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," 1997.
33. K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," 2014.