Sentiment Analysis of Electric Vehicles in India

Abstract:

The advent of technology has led to the emergence of electric vehicles. These eco-friendly modes of transportation have the potential to reduce pollution and have garnered much attention among consumers. In order to gain insight into consumer opinions and perceptions, it is essential for automotive sales companies to monitor relevant online forums. Using web scraping techniques, we collected data from popular Indian automotive websites and processed it using natural language processing (NLP). To analyse the data, we employed various deep learning techniques such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Simple Recurrent Neural Network (SimpleRNN), Bidirectional Long Short-Term Memory (BiLSTM), and CNN-LSTM. After comparing their accuracies, we found CNN-LSTM to be the most effective deep learning model, with an accuracy of 97.2%. This level of accuracy provides valuable insight into consumer sentiment and can be useful for automotive companies seeking to understand their target audience.

Keywords: Electric Vehicles, Automotive websites, Natural Language Processing, Convolutional Neural Network, Deep Learning.

1. Introduction:

Electric vehicles serve as a critical means to reduce pollution and provide an eco-friendly alternative to traditional petrol or diesel-based engine vehicles. In India, vehicles are responsible for contributing 20-30% of the air pollution (as stated by the International Energy Agency) [14]. Adopting the use of electric vehicles can significantly reduce this pollution. Additionally, non-renewable resources are rapidly depleting, and as a result, alternative options are necessary to meet our needs. Electric vehicles can serve as a viable alternative to traditional vehicles in India. In fact, the Indian government has recent approved a new $500 million Electric Vehicle (EV) Policy to promote India as a manufacturing hub for EVs and attract investment from global EV manufacturers [4].

Sentiment analysis is a process that involves analysing opinion data or views expressed in text to draw conclusions based on these opinions. In the case of electric vehicles, sentiment analysis can reflect user sentiment or opinion towards EVs, categorized as negative, positive, or neutral. Analysing user sentiment towards EVs can play a vital role in decision-making and manufacturing of EVs. People often share their opinions on automotive sites, providing reviews and comments that can prove beneficial for analysing user sentiment towards vehicles. Automotive websites serve as a direct and authentic source of user opinions. Deep learning techniques can be highly useful for sentiment analysis. Deep learning algorithms that analyse sentiment using deep neural network, such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), SimpleRNN can be useful. Hybrid models, such as CNN-LSTM, can also provide improved results [6].

2. Methodology:

This consists of data collection, data preprocessing, data analysis, training model, results, and conclusion. The following is the systematic research method used, represented in Figure 1.

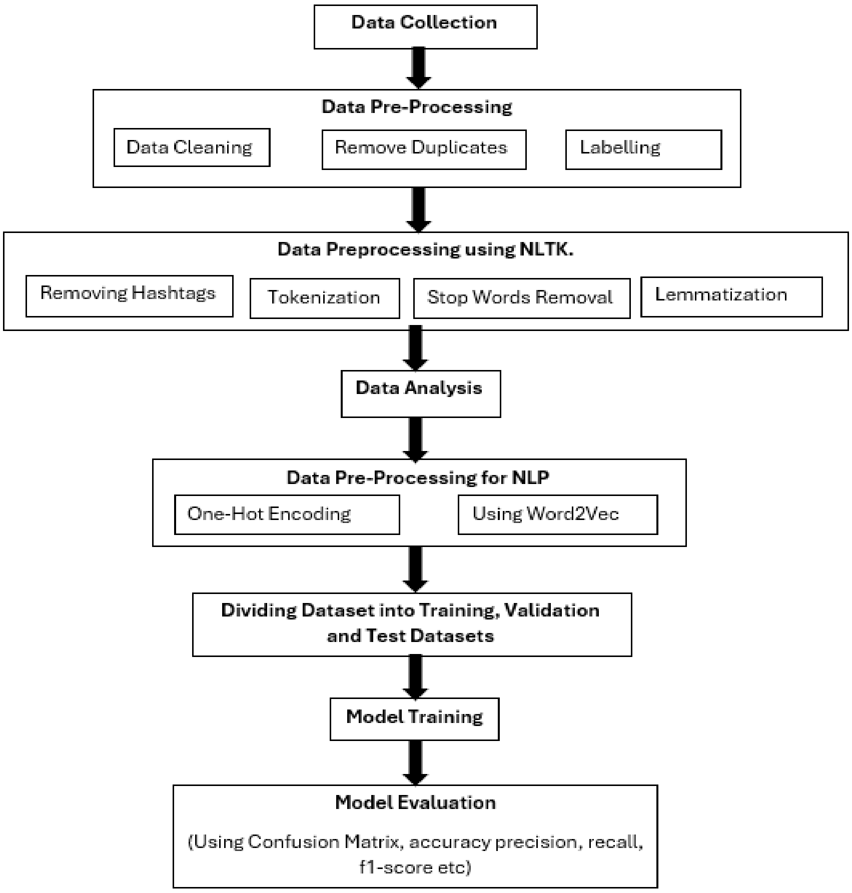


Figure 1. Research Method

3.1 Data Collection:

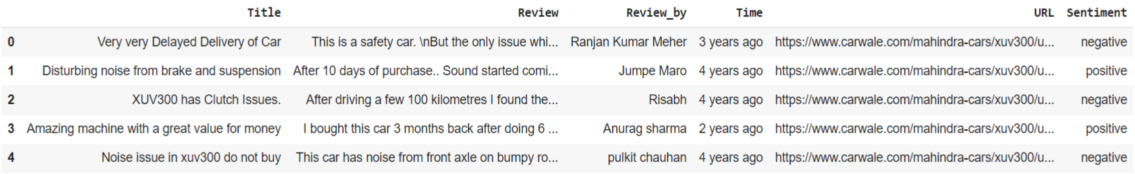
The methodology employed in this study involves several steps aimed at conducting sentiment analysis of electric vehicles (EVs) in India. Initially, data collection was performed by targeting some Indian automotive websites known for selling electric vehicles and also include reviews related to EVs. This selection included a mix of general automotive sites and those specializing in EVs. Websites such as Bikewale.com, Bikedekho.com, Cardekho.com, and Carwale.com were identified as primary sources for data collection. The web scraping technique was employed using Octoparse, a web scraping tool, to extract reviews and relevant content from these websites. The distribution of the data collected from different websites are as following represented in Table 1.

Table 1. Websites used for Data Collection

Websites	Reviews
Carwale.com	618
Bikedekho.com	875
Bikewale.com	43
Cardekho.com	571
Total	2107

3.2 Data Preprocessing:

Upon obtaining the raw data, preprocessing steps were undertaken to ensure its suitability for analysis. The cleaning phase involved removing irrelevant information such as HTML tags, punctuation, and stop words (common words like "the," "a"). Additionally, language processing techniques such as lemmatization were applied to standardize the text data by reducing words to their root forms. Then next step is to label manually as positive, negative and neutral on the basis of the sentiment. The following is an example of the labelled data as shown in Figure2.



	Title	Review	Review_by	Time	URL	Sentiment
0	Very very Delayed Delivery of Car	This is a safety car. \nBut the only issue whi...	Ranjan Kumar Meher	3 years ago	https://www.carwale.com/mahindra-cars/xuv300/u...	negative
1	Disturbing noise from brake and suspension	After 10 days of purchase.. Sound started comi...	Jumpe Maro	4 years ago	https://www.carwale.com/mahindra-cars/xuv300/u...	positive
2	XUV300 has Clutch Issues.	After driving a few 100 kilometres I found the...	Risabh	4 years ago	https://www.carwale.com/mahindra-cars/xuv300/u...	negative
3	Amazing machine with a great value for money	I bought this car 3 months back after doing 6 ...	Anurag sharma	2 years ago	https://www.carwale.com/mahindra-cars/xuv300/u...	positive
4	Noise issue in xuv300 do not buy	This car has noise from front axle on bumpy ro...	pulkit chauhan	4 years ago	https://www.carwale.com/mahindra-cars/xuv300/u...	negative

Figure 2. Labelled dataset

3.3 Data Analysis:

The dataset obtained after preprocessing consisted of 2107 reviews from various automotive websites. The structure of the dataset was analysed to understand the distribution of reviews across different platforms. Exploratory data analysis techniques were employed to gain insights into the characteristics of the data. The figure3 shows the amount of data based on label. The following figure 4 shows the sentiment over time among people.

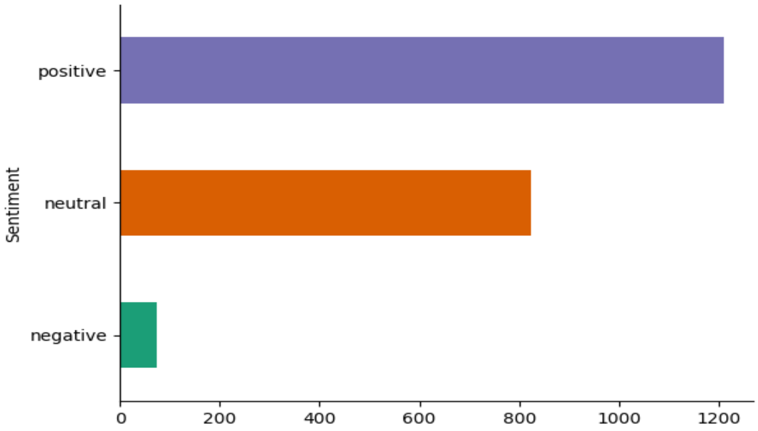


Figure 3 Amount of data based on label.

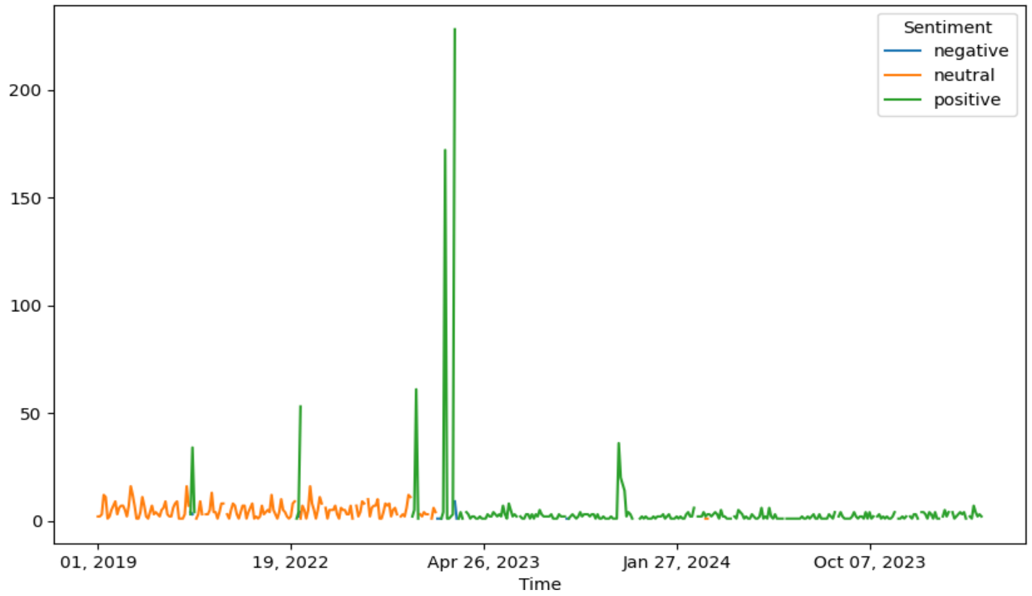


Figure 4. Sentiment Over Time

Word cloud is represented in figure 5. The word cloud analysis revealed that positive sentiments were mainly linked to environmental benefits, energy efficiency, and technological innovation. Conversely, negative sentiments were evident regarding concerns such as range anxiety, charging infrastructure, and upfront costs.

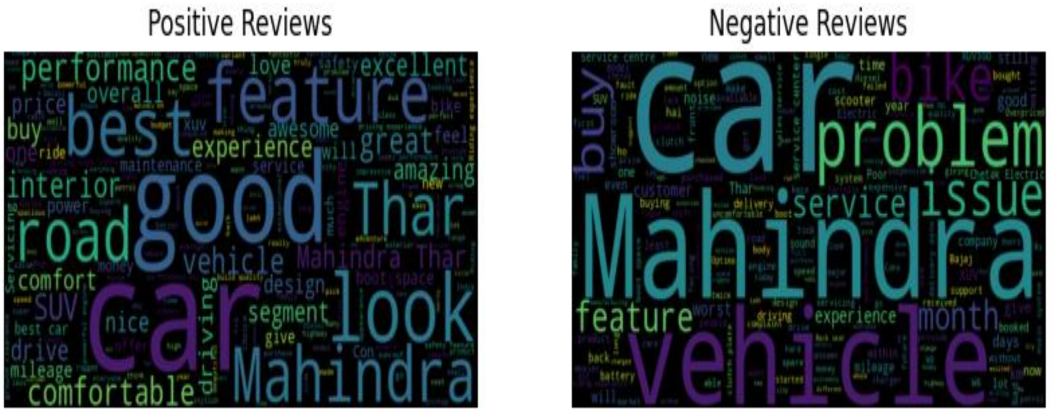


Figure. 5. Word Cloud Representation

The top positive and negative review in the given dataset as shown below in table 2.

Table 2. Top Reviews in Dataset

Top 5 Positive Reviews	**Top 5 Negative Reviews**
Best	Cost
Comfortable	Service
Good	Feature
Look	Charging
Excellent	Problem

3.4 One-hot Encoding:

For natural language processing (NLP), further preprocessing steps were undertaken to prepare the data for deep learning models. One-hot encoding was used to represent categorical data numerically. Additionally, word2vec, a technique for representing words as dense vectors in a continuous vector space, was employed for feature engineering.

3.5 Dividing the dataset into Training, Validation and Test:

Dataset is divided into the ratio of 0.7 of total as training and rest 0.15 for both testing and validation set. Total 2107 reviews are divided into following sets as shown in Table 3.

Table 3. Dataset Distribution into Training, Validation and Test Datasets

Dataset	**Size**
Training Set Size	1474
Validation Set Size	316
Testing Set Size	317

3.6 Deep Learning:

Deep learning models, including Long Short-Term Memory (LSTM), Simple RNN, Bidirectional LSTM (BiLSTM), Convolutional Neural Network (CNN), and CNN-LSTM, were chosen for sentiment analysis due to their ability to capture complex patterns in text data.

1. Long Short-Term Memory (LSTM): LSTMs are a type of recurrent neural network (RNN) specifically designed to address the vanishing gradient problem. They excel at handling sequential data with long-term dependencies, making them ideal for tasks like machine translation, speech recognition, and time series forecasting.

2. Simple Recurrent Neural Network (SimpleRNN): SimpleRNNs are the basic building block of RNNs. They process data sequentially, passing information from one layer to the next. However, they struggle with capturing long-term dependencies, limiting their effectiveness for complex tasks requiring historical context.

3. Bidirectional LSTM (BiLSTM): BiLSTMs are a variation of LSTMs that process data in both forward and backward directions. This allows them to capture context from both past and future elements in a sequence, leading to improved performance in tasks like sentiment analysis and text summarization.

4. Convolutional Neural Network (CNN): Though powerful for image recognition, CNNs are still being explored for sentiment analysis. They might be useful for extracting sentiment features from text data in a two-dimensional format (like word embeddings), potentially capturing local sentiment patterns within phrases.

5. CNN-LSTM: This hybrid model combines CNNs for extracting sentiment features and LSTMs for handling long-term dependencies in text sequences. It could offer a more comprehensive understanding of sentiment by capturing both local cues and contextual information, but further research is needed to compare its effectiveness to established models.

3. Results:

After preprocessing the data, which involved cleaning and language processing techniques, the dataset was structured for analysis.

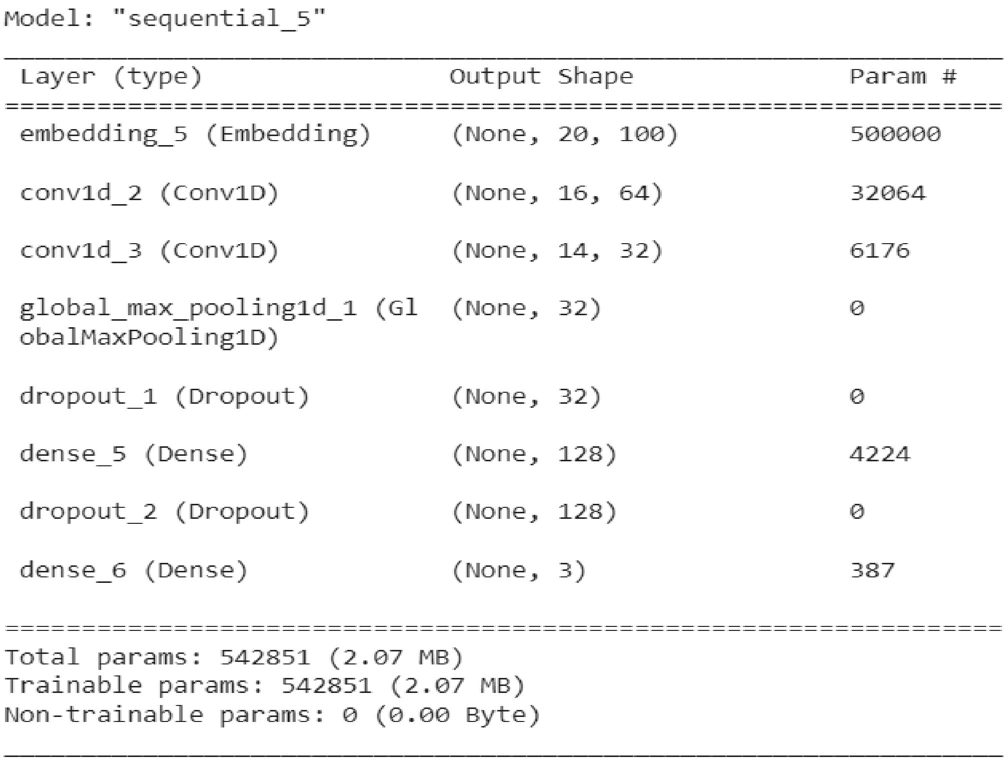
4.1 Model Performance:

Various deep learning models were evaluated for sentiment analysis. The performance of each model was assessed using metrics such as accuracy, precision, recall, and F1-score. Based on table 4, Among the deep learning models, CNN-LSTM exhibited the highest accuracies, with achieving an accuracy of 97.16%.

Table 4. Accuracies of different algorithms

Model	Accuracy
LSTM	96.85%
Bidirectional LSTM	95.58%
SimpleRNN	96.52%
CNN	96.84%
CNN-LSTM	97.16%

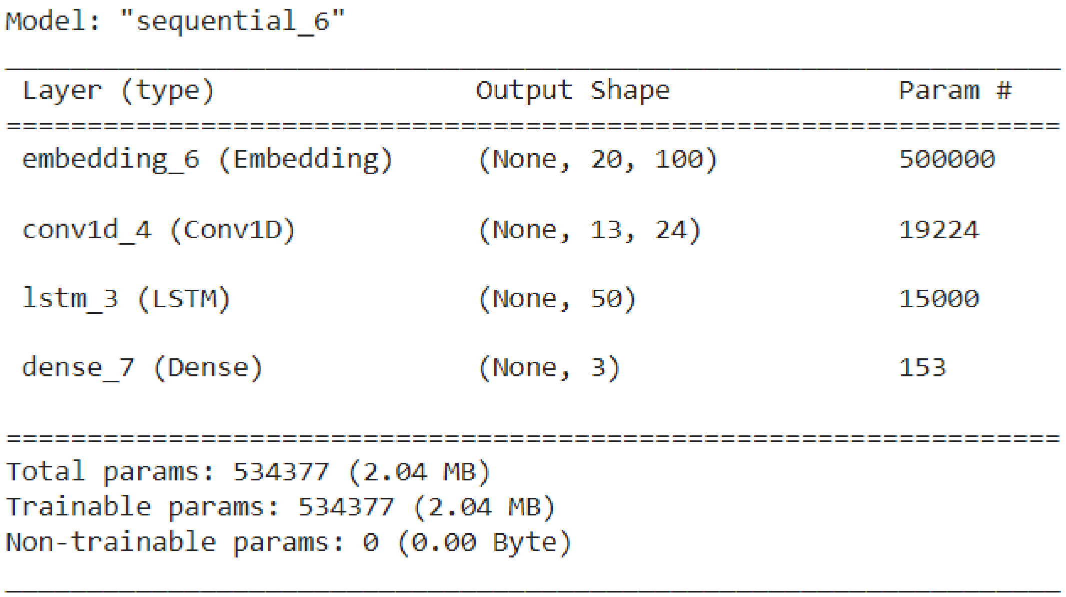
The architecture of the convolution neural network (CNN) as shown below in figure 6:



```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, 20, 100)           500000

 conv1d_2 (Conv1D)           (None, 16, 64)            32064

 conv1d_3 (Conv1D)           (None, 14, 32)            6176

 global_max_pooling1d_1 (Gl  (None, 32)                0
 obalMaxPooling1D)

 dropout_1 (Dropout)         (None, 32)                0

 dense_5 (Dense)             (None, 128)               4224

 dropout_2 (Dropout)         (None, 128)               0

 dense_6 (Dense)             (None, 3)                 387

=================================================================
Total params: 542851 (2.07 MB)
Trainable params: 542851 (2.07 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 6. Architecture of CNN (Convolution Neural Network)

The architecture of the best performing model (CNN-LSTM) as shown below in figure 7:



```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_6 (Embedding)     (None, 20, 100)           500000

 conv1d_4 (Conv1D)           (None, 13, 24)            19224

 lstm_3 (LSTM)               (None, 50)                15000

 dense_7 (Dense)             (None, 3)                 153

=================================================================
Total params: 534377 (2.04 MB)
Trainable params: 534377 (2.04 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 7. Architecture of CNN-LSTM (Convolution Neural Network-Long Short-Term Memory)

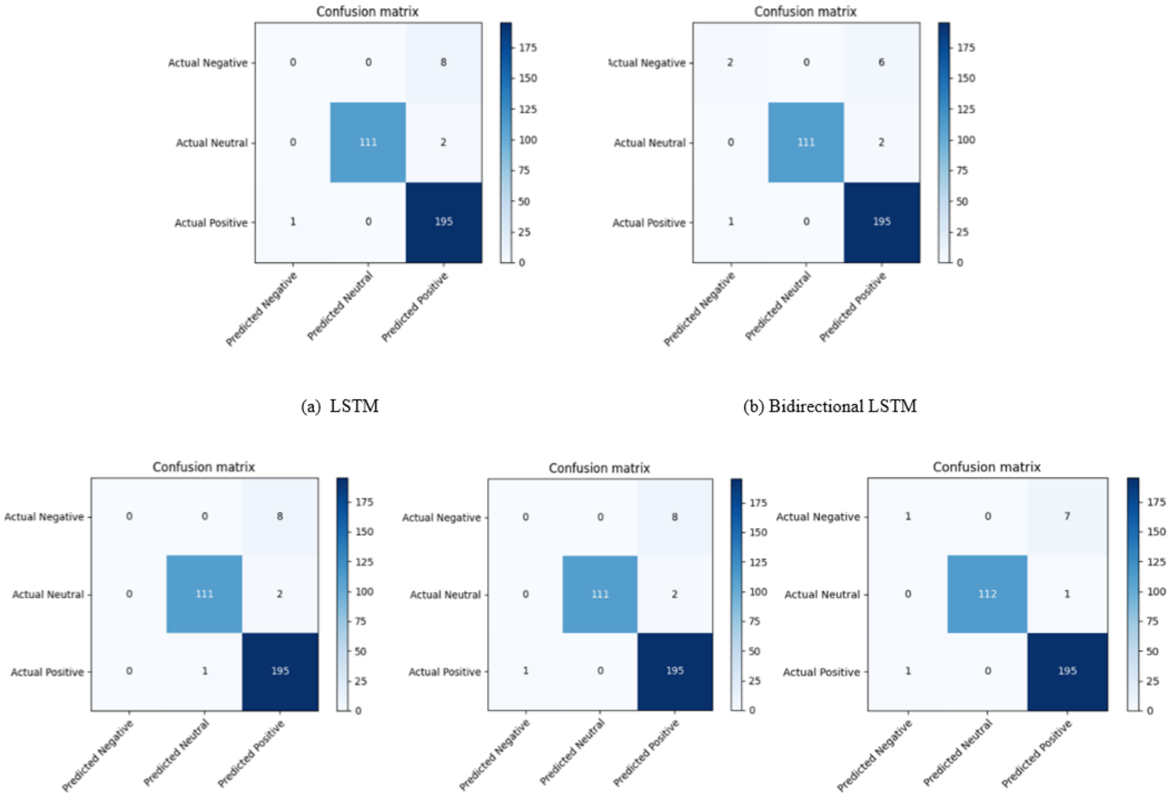
The performance of each model was further evaluated using precision, recall, and F1-score. CNN demonstrated high precision, recall, and F1-score, indicating its effectiveness in classifying sentiments accurately shown in table 5.

Table 5. Performance Metrics for Different Models

Model	Precision	Recall	F1-Score
LSTM	0.960	0.968	0.961
Bidirectional LSTM	0.960	0.955	0.957
SimpleRNN	0.940	0.970	0.950
CNN	0.970	0.970	0.970
CNN-LSTM	0.960	0.980	0.960

4.2 Confusion Matrix Analysis:

Confusion matrices were generated for each model to visualize the performance across different sentiment categories. The confusion matrices provided insights into the model's ability to correctly classify positive, negative, and neutral sentiments shown in figure 8.



(a) LSTM

(b) Bidirectional LSTM

(c) SimpleRNN

(d) CNN

(e) CNN-LSTM

Figure 8. Confusion Matrix

4. Conclusion:

In conclusion, our study on sentiment analysis of electric vehicles (EVs) in India highlights both the opportunities and challenges within the EV market. Positive sentiments surrounding environmental benefits and technological innovation are countered by concerns such as range anxiety and charging infrastructure. Our analysis of electric vehicle (EV) sentiment in India unveils a positive trend. Over time, we observed a significant shift from initial scepticism in 2019 to widespread positivity by 2023, underscoring the evolving sentiment surrounding EVs. Through deep learning techniques, particularly hybrid model CNN-LSTM(Convolutional Neural Networks- Long Short Term Memory), we achieved a high level of accuracy in sentiment classification.

Our findings offer actionable insights for automotive manufacturers and policymakers to address consumer needs and drive EV adoption. Future research efforts should focus on enhancing sentiment analysis models and exploring evolving consumer perceptions in the dynamic EV landscape. Additionally, the exploration of other advanced deep learning techniques holds promise for further advancing sentiment analysis in the EV domain.