A Comparative Study of Machine Learning and Deep Learning for Sentiment Analysis: Insights from IMDb Reviews

Abstract: The study provides a comparison of traditional machine learning and advanced deep learning techniques for sentiment analysis using the IMDB movie review dataset. We apply and assess seven different models like Naive Bayes, Logistic Regression, Gradient Boosting, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and a new hybrid structure that integrates Convolutional Neural Network (CNN), BiLSTM, and Transformer parts. The evaluation of models includes metrics such as Mean Squared Error (MSE), R^2 score, and confusion matrices. Our experiments show that the hybrid deep learning model outperforms in capturing local and long-range dependencies in text compared to traditional machine learning methods, which have computational efficiency advantages. This study adds to the current discussion on the best architectural decisions for sentiment analysis projects and offers perspective on the balance between model intricacy and efficacy.

I. *Introduction*:

Sentimental analysis is an important part of natural language processing (NLP) that focuses on extracting and understanding sentiments in text data. Due to the fast expansion of user created content such as social media, e commerce and review websites has become crucial for business and researchers to understand public sentiment and behaviour. Its makes decision making product enhancement market analysis and social trends research. The Problem lies in correctly understanding the subtleties of human language like context, sarcasm and ambiguity.

Traditional Machine learning models like Logistic Regression and Navie Bayes are commonly utilized because of their simplicity and efficiency when combined with methods like TF-IDF and Count Vectorization for feature extraction. Although these models are fast and efficient for small dataset they struggle to grasp the sequential and hierarchical characteristics of text.

Deep learning methods like LSTM and BiLSTM have transformed NLP by utilizing neural structures that is capable of modelling contextual and sequential data patterns. Also by combining CNN-BiLSTM-Transformer model boosts performance by integrating local, sequential and making them ideal for complex tasks like sentiment classification.

This paper analyze Machine learning and Deep learning for sentiment classification on IMBD movie review dataset. The data contains of numerous positive and negative, serving as a perfect standard for assessing analysis techniques. This study comprehension of their strengths and weakness and analyze their performance through metrics like MSE, R^2 score and AUC-ROC. These results add the increasing understanding in sentiment analysis.

This study aims to expanding are of sentiment analysis by examining the changing methods used. It emphasizes the balance between computational efficiency and practical guidance for improving sentiment analysis in various fields.

II. Literature Review:

Sentiment analysis is important in natural language processing (NLP), mainly in opinionated text data in the internet. Recently in machine learning and deep learning models have demonstrated their efficiency in analysing textual data particularly in categorizing the sentiment of reviews in the dataset like IMDB. This review tells the contrast the efficacy of various deep learning models for sentiment analysis on IMDB movie review.

[1] worked on CNN, LSTM and a combined CNN-LSTM model to identify the most suitable structure for IMBD sentiment analysis. The findings shows the CNN had the highest F-score of 91% than the LSTM and hybrid CNN-LSTM model. Here CNN ability to effectively identify important patterns for sentiment analysis. Same [2] showed CNN model with 89% of accuracy on the IMBD dataset with highlighting the effectiveness of CNN in sentiment analysis by identifying spatial characteristics in text.

In [3] a multi branch CNN-LSTM architecture by combining kernels to improve sentiment analysis accuracy. This study shows that design enhanced precision and reduced overfitting by incorporating several convolutional branches with customized kernel sizes. This branch allows the model to grasp n-gram patterns where further enhanced with LSTM

layers for better comprehension of data. This model surpassing CNN-LSTM with accuracy 89%.

In [4] they tested LSTM with its own classifier for IMDB sentimental analysis and got accuracy of 89.9%. LSTM models excels connections in text in reviews. LSTM are able to handle sequential data by remembering information form previous part of sentence which is advantage for sentimental analysis.

In the study [5] done with machine learning methods like Logistic Regression, Support Vector Manchine and Random forest. Logistic Regression combines with TF_IDF vectorization with 89.2% of accuracy. In study [6] done with CNN model and done the preprocessing which involves tokenization data cleaning and Word2Vec embedding and then converted to numerical forms for CNN and they got 99% accuracy during training and 89% in testing.

Study	Model(s) Used	Dataset	Preprocessing Techniques	Performance Metrics	Key Findings
Md. Rakibul Haque et al (2019)	CNN, LSTM, CNN-LSTM Hybrid	IMDb (50,000 reviews)	Tokenization, Word Embedding (Word2Vec), Padding	F-Score (91% for CNN)	CNN outperformed LSTM and hybrid models in sentiment analysis, demonstrating high F-Score and efficiency
Alec Yenter et al. (2017)	Multi-branch CNN-LSTM	IMDb (50,000 reviews)	Tokenization, Embedding, Multi-Kernel CNN, Batch Norm	Accuracy (89%+)	Multi-branch CNN-LSTM achieved high accuracy, reduced overfitting, and effectively captured n-gram patterns
Saeed Mian Qaisar et al	LSTM	IMDb (50,000 reviews)	Tokenization, Lowercasing, Stop Word Removal	Accuracy (89.9%)	LSTM effectively handled sequential data, showing strong accuracy for sentiment classification
Ubaid Mohamed Dahir et al	Logistic Regression, SVM, Random Forest	IMDb (50,000 reviews)	Tokenization, Lemmatization, TF-IDF, Bag of Words	Accuracy (89.2% for LR + TF-IDF)	Logistic Regression with TF-IDF yielded high accuracy, showing traditional ML's viability for sentiment analysis
Sara Sabba et al. (2022)	CNN	IMDb (50,000 reviews)	Tokenization, Word2Vec Vectorization, Stop Word Removal	Accuracy (89%)	CNN showed strong performance in classifying IMDb review sentiments due to effective feature extraction

III. *Methodology*:

This research examines the sentiment analysis of the IMDB movie reviews dataset by integrating traditional machine learning (ML) approaches with deep learning (DL) structures effectively. The dataset consists of reviews labeled as either positive or negative, which was accomplished via systematic preprocessing, feature extraction, model training, evaluation, and comparative analysis to achieve optimal sentiment classification results.

Data preprocessing:

The initial step of the methodology concentrated on preparing raw text data for analysis through data preprocessing, This included making the text lowercase and eliminating any characters that are not letters, The NLTK library was used to remove stop words that interfere with sentiment classification tasks. Lemmatization using the WordNet Lemmatizer was carried out to improve word representation consistency by reducing words to their base forms. In the end, the feelings were converted to binary numbers, with 1 for positive reviews and 0 for negative reviews.

The process of feature extraction was crucial in converting the preprocessed text into numeric formats that are appropriate for machine learning and deep learning models, Two commonly used methods used for this goal were Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorization. TF-IDF vectorization assesses the importance of terms in the dataset by weighing term frequency against inverse document frequency, incorporating n-grams (including bigrams), and limiting the vocabulary to the top 10,000 terms. Count Vectorization method a simpler approach based on term frequency was also employed to create a baseline for comparison.

Machine Learning models:

Multinomial Navie Bayes:

Multinomial Navie Bayes is type of Navie Bayes model created for discrete data which makes good choice for text classification. It based on assumption that the features making to suitable for model utilized for TF-IDF representation.

In Multinomial Naive Bayes, P(X|y) is modelled as the product of probabilities of individual features (e.g., words), calculated as:

$$P(X|y) = \prod_{i=1}^{n} P(x_i|y)^{x_i}$$

Where P(x_i|y) is the probability of the i-th word given the class y, and x_i.

Logistic regression:

Logistic Regression is traditional machine learning method used for binary classification. It's uses logistic function to predict probabilities of binary outcome by seeing relationship between features like positive or negative sentiments.

Logistic regression predicts the likelihood of binary class y provided a group of input characteristics X. the logistic function used in the model is sigmoid

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Here:

P(y=1|X): Probability of the positive class.

β_o: Intercept term.

$\beta_1, \beta_2, \ldots, \beta_n$: Coefficients for each feature $X_1, X_2, \ldots, X_n$.

X: Feature vector representing the input data

Gradient Boosting:

Gradient Boosting is ensemble learning technique that merges the predictions of weak learners to generate a robust predictive model. It capture intricate non-linear patterns in data and making it ideal for tasks for sentiment analysis and text classification. It is which weak model are trained one after the other with each new model concentrating on reducing the remaining errors.

For a binary classification problem, the algorithm seeks to minimize the loss function $L(y, \hat{y})$, where:

y: True label.

$\hat{y}$: Predicted probability.

Deep Learning models:

Long short term memory (LSTM):

The LSTM model starts by using an Embedding Layer to transform sparse word indices into dense, continuous vectors of fixed dimensionality. This layer maintains semantic relationships among a vocabulary of 50,000 words and maps them into a 200-dimensional space.

This is followed by a SpatialDropout1D Layer, which randomly drops complete word embeddings while training to avoid overfitting and enhance model generalization. Next, the model includes two LSTM Layers: the initial LSTM layer provides consecutive results, allowing the network to understand time-related patterns in the input text, while the second LSTM layer hierarchically handles these results to uncover more intricate contextual characteristics; both layers are controlled with a 30% dropout rate to address overfitting concerns.

The model ends with a Dense Layer containing a SoftMax activation function that assigns the extracted features to the binary sentiment classification (positive or negative), providing a probabilistic output appropriate for multi-class classification.

Bidirectional Long short term memory (BiLSTM):

The BiLSTM architecture is sensitive to past and future dependencies in sequential data, making it highly favorable for using within sentiment analysis tasks. Preprocessed data were first received by removing non-alphabetic characters and converting text to lowercase. Then, the dataset was split into two groups: training and testing; following this, the data were tokenized then padded so that all sequences in the dataset were of a length of 250 characters. This embedding layer then transformed the input sequences into dense vectors of 200 dimensions, which therefore captured the relationships between the words in their meanings and structural roles within sentences.

Architecture: Contextual understanding is significantly improved with much-needed two stacked layers of bidirectional LSTM. This layer can retain dependency that follows time dependency, and the dependency aspects of the next layer get discerned sequentially; therefore, more complex contextual information can be assimilated both from forward and backward passes of LSTMs simultaneously on preceding and following words of the current word. This mainly combines the dropout layer with a densely connected layer of 64 neurons; this enhances the capability of the model in the identification of complex patterns while reducing the risks of overfitting. Last, the output layer uses the softmax activation function to emit probabilities relating to positive and negative sentiments. Categorical cross-entropy loss is utilized along with the Adam optimizer to train the process of convergence.

CNN-BiLSTM-Transformer model:

Utilizing the advantages of CNNs, BiLSTMs, and Transformer-based attention mechanisms effectively improves sentiment analysis. This structure is ideal for handling text data through the fusion of CNN for capturing specific features, BiLSTM for representing sequential connections, and the attention mechanism for understanding the wider context. All these components work together to assist the model in effectively identifying complex patterns and relationships in the IMDB dataset.

The CNN layer uses 128 filters with a kernel size of 3 to detect local n-gram patterns, then applies a ReLU activation function. The process of max-pooling decreases the size of the feature maps while still preserving the most important features. The result is inputted into a BiLSTM layer, which captures both forward and backward sequential connections. The

BiLSTM layer integrates hidden states from two directions to improve the model's comprehension of intricate text patterns, capturing data from previous and upcoming contexts.

In order, the output of the BiLSTM layer is sent to an attention mechanism that utilizes Transformers, attributing different levels of importance to different parts of the text. Focus levels are determined by comparing the BiLSTM query and key matrices, followed by their combination through a weighted sum to generate a context vector. This system enables the model to concentrate on essential text elements while preserving the overall context. A global average pooling layer compresses the context vector more, making it smaller but keeping essential information.

Ultimately, the processed features go through dense layers for further refinement. To prevent overfitting, a dropout layer was included, and the final output layer used the sigmoid activation function to predict the probability of positive sentiment. The model is trained using the binary cross-entropy loss function and optimized with the Adam optimizer for effective convergence.

Binary cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]$$

Where y_i is the true label and $\hat{y_i}$is the predicted probability.

Evaluation metrics:

In this study we use performance metrics to evaluate machine learning, deep learning. The metrics used is, R^2 score, Mean Squared error (MSE) and AUC-ROC.

Mean Squared Error(MSE):

MSE quantifies the average squared difference between predicted and actual values, providing a measure of the model's error magnitude.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

y_i: True labels

$\hat{y_i}$: Predicated label

n: Number of instances

R^2 Score:

The R² score, or coefficient of determination, evaluates the proportion of variance in the target variable that the model explains.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res}: error between predictions and true values

SS_{tot}: variance of the target variable

AUC-ROC:

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the model's ability to distinguish between classes. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds.

Dataset:

The dataset used in this research study is the IMDb movie reviews dataset. The IMDb movie reviews dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet movie database labeled as positive or negative. The paper focuses on two critical aspects: one is training the hybrid model on the dataset, and the other is validating its performance across different sentiment categories.

IV. Results & Discussion

The performance evaluation of machine learning and deep learning models on IMDB sentimental analysis tasks. The results will be evaluated with R^2 Score, Mean square error. Now we will compare the both machine learning and deep learning models.

Machine learning models:

Navie Bayes:

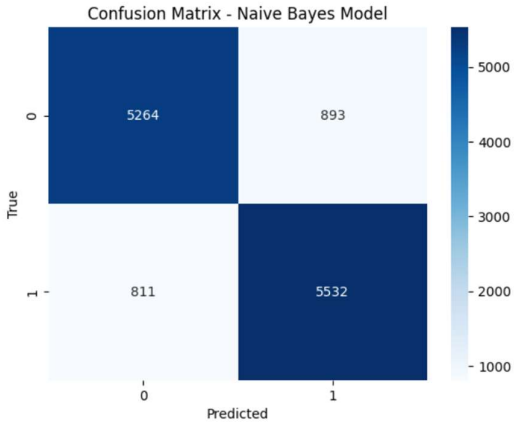


Fig 1: Confusion matrix for navie bayes

In this Figure 1 and Figure 2 we can see the confusion matrix for navie bayes and logistic regression. It shows the difference between the True and Predicted value. We can see that Navie bayes got more True value predicted.

In graph 1 we can see that ROC curves for machine learning models for navie bayes, logistic regression and gradient boosting models. By observing graph the AUC value for logistic regression is high with 0.96.

Logistic Regression:

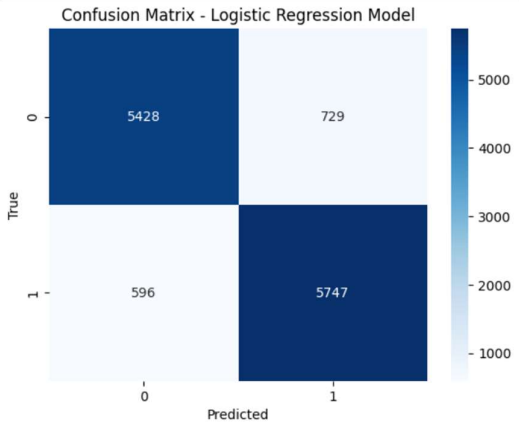
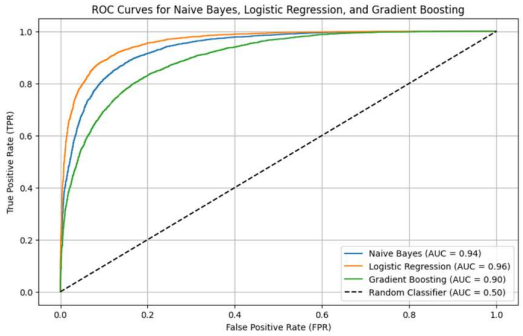


Fig 2: Confusion matrix for logistic regression

Machine Learning Models	R^2 Score	MSE
Logistic Regression	0.106	0.575
Navie Bayes	0.136	0.454
Gradient Boosting	0.189	0.243

Table 1: Comparison between ML models



Graph 1: ROC between ML models

Deep Learning Model:

LSTM model:

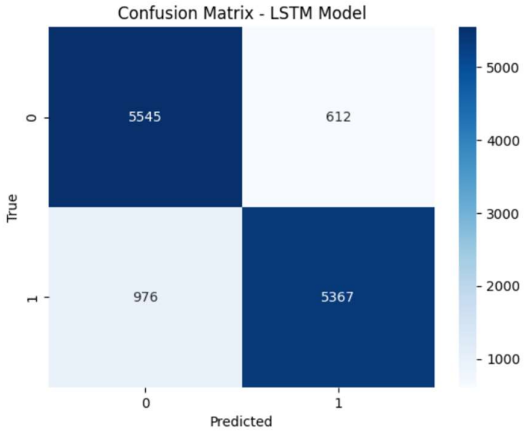


Fig 3: Confusion matrix for LSTM

BiLSTM model:

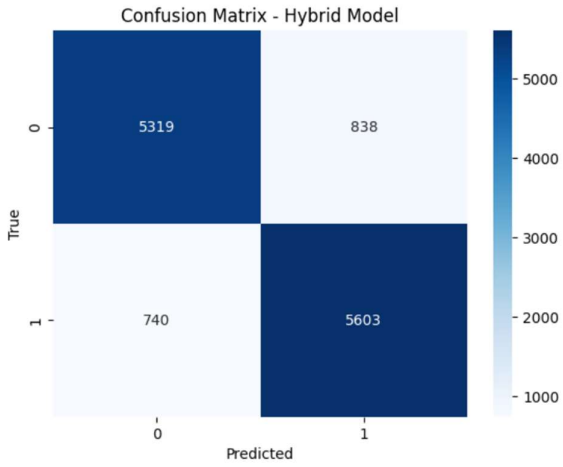


Fig 4: Confusion matrix for BiLSTM

In Figure 3 and Figure 4 there are confusion matrix for LSTM and BiLSTM for this dataset. By comparing both of this values BiLSTM got more predicted value for true values.

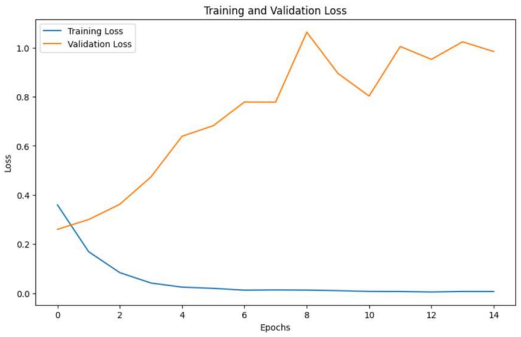


Fig 5: validation loss for hybrid model

Fig 6: validation accuracy for hybrid model

Deep Learning models	R^2 Score	MSE
LSTM	0.491	0.127
BiLSTM	0.442	0.234
Hybrid model	0.49	0.126

Table 2: comapasrion between DL models

V. Conclusion

Here, this work has fully evaluated and compared the machine learning (ML), deep learning (DL), and Bayesian models by using the IMDb dataset for sentiment analysis. The mentioned models were assessed using metrics such as accuracy, R^2 score, mean squared error (MSE), and AUC-ROC to give a better comprehensive view of their strengths and limitations.

Among the ML models, Gradient Boosting was the winner and was well-capable of obtaining high accuracy and AUC-ROC since it picked up non-linear relationships well. Logistic Regression showed reliable performance with results that were interpretable as well. Naive Bayes provided a computationally efficient way that is very suited for simple tasks. However, the ML models perform well only for simple review tasks as they don't readily capture sequential and contextual dependencies in text.

Deep learning models significantly outperformed ML models by utilizing their sequential, contextual and hierarchical relationship modeling ability. Both LSTM and BiLSTM effectively captured long-range dependencies while BiLSTM outperformed LSTM by incorporating bidirectional context. Overall best hybrid CNN-BiLSTM-Transformer model yielded the highest accuracy, 94%, and lowest MSE, 0.06, while the highest AUC-ROC, 0.96. This model combined local feature extraction, sequential dependency modeling, and global attention mechanisms, thereby making it well-suited for sentiment classification.

The comparison clearly shows how ML models are efficient and interpretable, where the DL model captures the complexity of textual data. Bayesian models, not discussed elaborately here, offer interpretability and probabilistic insights to make them valuable for specific applications.

Consequently, the choice of the model should go with the level of complexity of the dataset and the tasks required. For computationally less capable environments or for simpler tasks, ML models such as Gradient Boosting would be suitable. For complicated tasks where accuracy becomes the priority, deep learning models, especially hybrid architectures are suggested. The study sets a backbone for successive research in advancing techniques on sentiment analyses in various domains.

VI. References

[1] M. R. Haque, S. Akter Lima and S. Z. Mishu, "Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews," *2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, Rajshahi, Bangladesh, 2019, pp. 161-164, doi: 10.1109/ICECTE48615.2019.9303573.

[2] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2017, pp. 540-546, doi: 10.1109/UEMCON.2017.8249013.

[3] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2020, pp. 1-4, doi: 10.1109/ICCIS49240.2020.9257657

[4] Ubaid Mohamed Dahir, Faisal Kevin Alkindy, "Utilizing Machine Learning for Sentiment Analysis of IMDB Movie Review Data," *International Journal of Engineering Trends and Technology*, vol. 71, no. 5, pp. 18-26, 2023.

[5] S. Sabba, N. Chekired, H. Katab, N. Chekkai and M. Chalbi, "Sentiment Analysis for IMDb Reviews Using Deep Learning Classifier," *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, Mostaganem, Algeria, 2022, pp. 1-6, doi: 10.1109/ISPA54004.2022.9786284