Exploring Hybrid Approaches for Sentiment Classification: A Comparative Study of LSTM, Naive Bayes, and Bayesian Network on IMDB Reviews

1st Koduru Balavignesh Reddy
dept.Computer Science,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Amaravati, India
av.en.u4aie22016@av.students.amrita.edu

2ndDokala Sai Parthava Naidu
dept.Computer Science,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Amaravati, India
av.en.u4aie22010@av.students.amrita.edu

3rdNallani Deekshitha
dept.Computer Science,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Amaravati, India
av.en.u4aie22024@av.students.amrita.edu

4th Dr. M. Srinivas
dept.Computer Science
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Amaravati, India
msrinivas@av.amrita.edu

***Abstract*—This paper evaluates the use of different types of models for machine learning and deep learning applied to sentiment analysis on the IMDb movie reviews dataset. Classical approaches, including Naive Bayes, Logistic Regression, and XGBoost, are compared with sequential deep architectures like (LSTM) and Recurrent Neural Networks (RNN). A Bayesian network is added to the portfolio in order to provide probabilistic inference for sentiment prediction. Techniques applied to preprocess the data include BoW and tokenization with padding. Results show that LSTM and RNN significantly identify sequential text patterns and perform well at high accuracy and F1 scores especially if input sequences are tokenized and padded. Classical models like Naive Bayes and Logistic Regression achieved competitive precision and recall capabilities with good computational efficiency and interpretability by using BoW features. XGBoost proved to be robust with optimized gradient-boosting techniques, while the Bayesian network presented a novel probabilistic approach, enabling the system to provide a probability of sentiment prediction based on the relationship between variables of sentiment and prediction. Results bring to light the potential integration of deep models with the capabilities of sequential learning and classical approaches or probabilistic models such as Bayesian networks for developing a scalable and efficient sentiment analysis framework for real-world applications. The experimental setup involves using LSTM, XGBoost, Naive Bayes, Logistic Regression and Bayesian Networks models to analyze IMDB movie reviews for sentiment classification has achieved 85.25, 86.13, 83, 90.21 and 80 accuracies respectively.**

***Index Terms*—Sentiment Analysis, Naive Bayes, Logistic Regression, XGBoost, LSTM, Bayesian Network, Bag-of- Words, Feature Extraction.**

979-8-3315-2751-8/25/$31.00 ©2025 IEEE

I. Introduction

Sentiment analysis is actually the computational task of identifying and extracting subjective information from text. It has become a fundamental instrument in Natural Language Processing (NLP). When examining online contents, especially through user generated texts such as movie reviews, social media posts, and feedback on various platforms, there is an exponential increase in the volume of content. Sentiment analysis will play a crucial role in formulating valued insights from this unruly giant pool of unstructured data. The ability to interpret public sentiment influences the decision and business strategies in plenty of industries, especially in the entertainment and marketing sectors as well as customer service. Hence, the IMDb movie reviews dataset would be a great benchmark for the analysis, as it is abundant labeled text data with binary sentiment classification, positive or negative, which may be explored with different approaches through machine learning and deep learning. It has long been the use of the traditional machine learning algorithms including Naive Bayes, Logistic Regression, and XGBoost for

accomplishing text classification tasks, offering reliable and interpretable models. These days, though the complexity of the text increases, sequential models, which embrace the use of Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), are becoming popular. These models capture contextual dependencies and patterns in text, which would allow for an actual deeper understanding of the sentiment embedded in words and phrases.

In parallel, probabilistic methods such as Bayesian networks serve as yet another layer of interpretation while modeling dependencies and conditional probabilities among variables, suggesting a different view on the task of sentiment prediction. Classical machine learning algorithms, despite their simplicity and interpretability, are still highly relevant in many real-world applications where deep learning models outperform them due to their capacity for capturing intricate relationships in text data.

It will compare such diverse techniques on the IMDb movie reviews dataset, explore their strengths and weaknesses, and determine which approach best suits this sentiment classification task with regard to precision, recall, accuracy, F1-score, and interpretability. Another objective explored in this comparative analysis is that of combining different models-leveraging the sequential capabilities of deep learning with the efficiency and interpretability of classical machine learning and probabilistic models-building a scalable and robust sentiment analysis system for large-scale text data.

II. LITERATURE REVIEW

In this section, we discuss various research papers that apply machine learning algorithms to sentiment analysis on IMDB movie reviews, emphasizing model accuracy, algorithm choice, and evaluation metrics. Anwar Ur Rehman et al. [1] study combines CNN and LSTM models for IMDB movie review sentiment analysis. CNN captures local textual features, while LSTM processes sequential dependencies. This hybrid approach improves accuracy, leveraging CNN's feature extraction and LSTM's sequence understanding. The study emphasizes ensemble techniques for robust sentiment classification. Abhijaat Pandey et al. [2] research evaluates sentiment analysis methods on IMDB reviews using classical machine learning and deep learning models. Techniques like Support Vector Machines (SVM) and transformers are compared. The study highlights transformers for their superior performance, although traditional models remain efficient for smaller datasets. Wang and Li [3] use lexicon-based approaches alongside neural networks for sentiment classification. Lexicon techniques efficiently preprocess data, while neural networks capture complex patterns. The combination balances computational efficiency and accuracy, emphasizing neural networks' adaptability to textual nuances.

Beatriz Alejandra Bosques Palomo et al. [4] work explores deep learning techniques like GRU and CNN for IMDB sentiment analysis. GRU excels in handling long-term dependencies, while CNN efficiently extracts features. The study achieves competitive results by integrating these architectures, showcasing the adaptability of deep learning to text data. Neelisetty Sri Lakshmi Sai Charitha et al. [5] compared Naive Bayes, Decision Trees, and LSTM, emphasizing LSTM's superiority in sequential tasks while noting the interpretability of Naive Bayes. Ashwin Kumar [6] examined RNN and CNN models, recommending hybrid approaches to utilize RNN's sequential capabilities and CNN's feature extraction strengths. Meta Mahyarani et al. [7]implemented Naive Bayes with Information Gain for feature selection, enhancing IMDB sentiment classification. Information Gain improves feature relevance, leading to better precision and recall. This combination highlights the efficiency of probabilistic models in structured datasets.

Saeed Mian Qaisar [8] focused on LSTM, demonstrating its suitability for capturing long-term dependencies in IMDB reviews with high accuracy. K. Amulya et al. [9] comparative analysis evaluates machine learning and deep learning models, including Random Forest and LSTM, for IMDB sentiment analysis. LSTM demonstrates superior accuracy in handling unstructured data, while Random Forest excels in interpretability, emphasizing the trade-offs between performance and explainability. Sandesh Tripathi and Ritu Mehrotra [10] study applies RNN and GRU models to IMDB datasets. These recurrent architectures handle sequential dependencies effectively, achieving up to 90 percent accuracy. GRU is preferred for its computational efficiency compared to standard RNNs, showcasing its potential in resource-constrained environments. Puspasourav Panda et al. [11] analyzed university and IMDB reviews, identifying hybrid approaches as solutions for sarcasm detection and data quality issues. Harish et al. hybrid feature extraction method combines TF-IDF and Word2Vec embeddings for IMDB sentiment analysis. This integration captures semantic and frequency-based features, improving model accuracy. The study underlines hybrid techniques' significance in extracting diverse textual features. Sara Sabba et al. [12] work uses deep learning classifiers, including BiLSTM and CNN, for sentiment classification. BiLSTM captures bidirectional dependencies, and CNN extracts hierarchical features. The combination yields high accuracy, demonstrating the advantages of deep architectures in complex textual datasets. Qianwen Ariel Xu et al. [13] systematically review sentiment analysis on social media data, focusing on lexicon-based and deep learning approaches. They highlight emerging trends like transformers for nuanced sentiment detection and challenges in handling multilingual datasets and contextual nuances. Zeeshan Shaukat et al. [14] integrated lexicon-based features with neural networks, achieving high precision and showcasing the benefits of combining traditional and advanced approaches.

III. METHODOLOGY

Testing three models of sentiment analysis upon 50,000 IMDb movie reviews which were labeled positive or negative is the methodology of this study. The first model used was LSTM, where the neural network was trained on a padded sequence of tokenized text data. The embedding layers based LSTM was evaluated against accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The second model, Naive Bayes, extracted the features using the technique called the Bag of Words (BoW) with CountVectorizer, and its performance was also measured in terms of accuracy, precision, recall, F1-score, confusion matrix, and ROC- AUC. In this line, the third model was XGBoost, where it also extracted the features using Bag of Words and evaluated in the same evaluation metrics. All the models were compared based on such factors to assess which method is the best to apply for sentiment classification on the dataset.

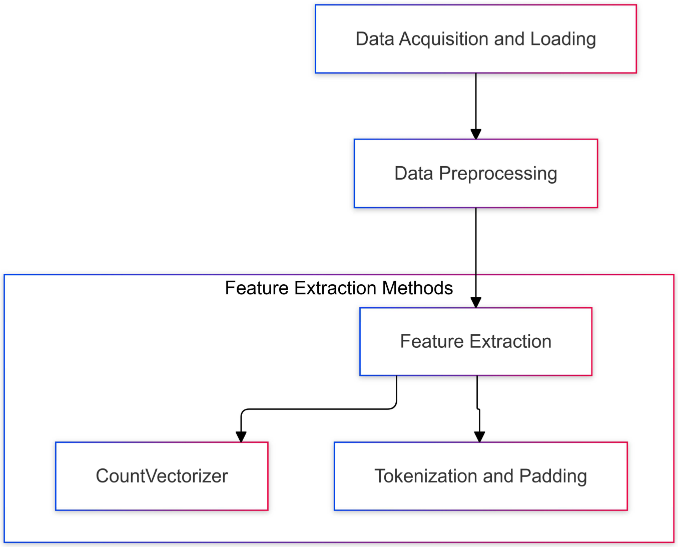


Fig. 1. Data Preparation

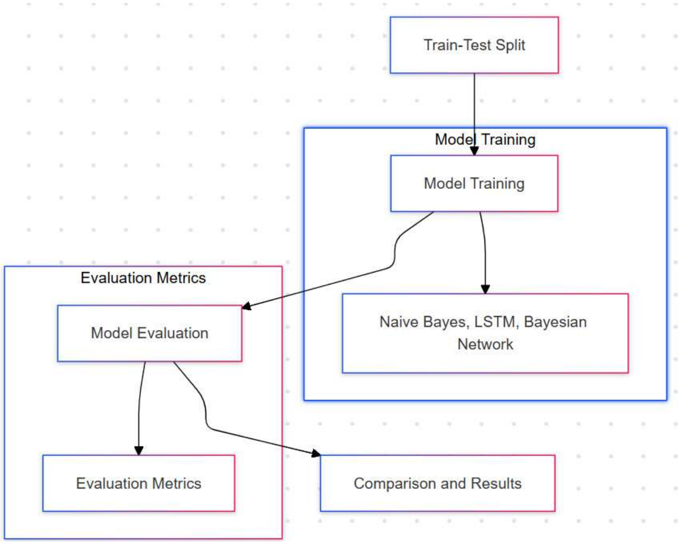


Fig. 2. Model Training and Evaluation Workflow

A. Dataset Description

The dataset for this experiment consists of 50,000 movie reviews collected from the IMDb platform. There are 25,000 and 25,000 positive and negative reviews respectively, so there is no class imbalance at all. Each text comes in the form of unstructured data, so preprocessing can include tokenization, padding, or vectorization in order to be accepted by the model. This dataset can be used as a benchmark in sentiment analysis tasks; it can be split into training for 60 percent. This is to ensure that the model has enough data to make training and thereby several performance evaluations.

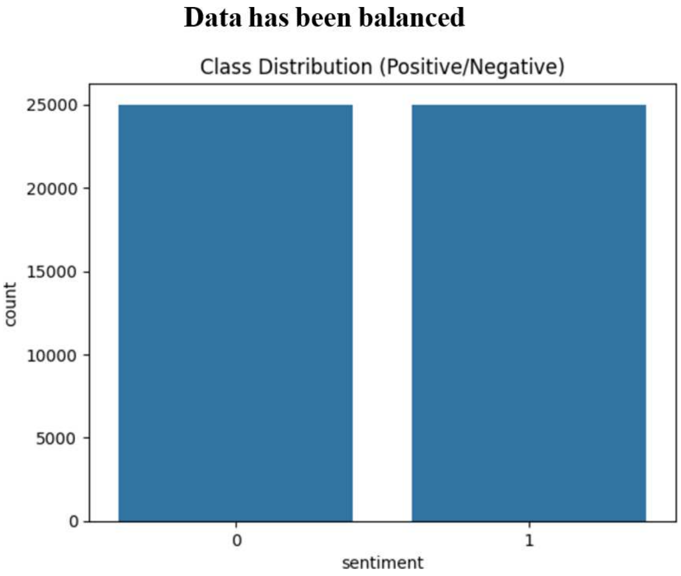


Fig. 3. Class Distribution

B. Feature Extraction

Feature extraction is very fundamental in the preparatory steps of ml and dl on textual data. There were two entirely different approaches used in this study for feature extraction. Bag of Words (Bow) was used in the implementation of classical machine learning models, Naive Bayes, Logistic Regression, and XGBoost. Text data was converted into numerical feature vectors through the ‘Count Vectorizer‘. This captures the word frequencies across the reviews. The process produced a sparse matrix representation of text with a maximum of 5000 features. Techniques of tokenization and padding were used for LSTM and RNN-based deep learning models. Text data was tokenized using the class ‘Tokenizer‘ that transforms a text sentence into an integer-sequence. The size of vocabulary was restricted to the 5000 most occurring words, and sequences were padded or truncated to a constant length with 200 tokens through ‘pad sequences‘ in such a way that input size is similar for samples. This technique lets the sequential models gain adequate ability to classify temporal patterns in the text data, hence suitable for sentiment analysis tasks.

C. Model Selection

The article used four different models to perform sentiment analysis on IMDb movie reviews and illustrated various techniques applicable for machine learning and deep learning techniques. A Naive Bayes classifier, based on Bayes’ theorem, was applied using Bag of Words as the feature extraction that transforms text data into numerical vectors using word counts. It is computationally less expensive and appropriate for classification with binary choices. Logistic Regression is a very popular linear model, and the Bow feature was considered to utilize the model’s simplicity and effectiveness for binary classification. XGBoost will bring in some gradient-boosting algorithm techniques that improves the accuracy of predictions with optimized boosting. Hence, XGBoost will be useful for text-based jobs. Finally, a Long Short-Term Memory neural network was used to capture sequential dependencies in text data by using tokenized and padded sequences to model the temporal structure of language. One model was selected based on its strength in handling sentiment analysis, thereby providing a comprehensive evaluation across different approaches.

Naive Bayes: Naive Bayes is a simple probabilistic classifier that applies Bayes’ theorem under strong (naive) independence assumptions between the features. This model is particularly well-suited for text classification tasks since the independence assumption holds quite well in practice. The model is trained on the Bow features, and the variant used here for Naive Bayes is MultinomialNB, as it better suits word frequency data.

Logistic Regression: Logistic Regression is a linear model designed for the problem of a binary classification. It estimates the probability of a binary outcome in relation to the linear combination of input features. In this paper, the Bow features are fed into Logistic Regression aiming to predict whether a review carries a positive or negative sentiment.

XGBoost: XGBoost stands for extreme gradient boosting. It’s a highly scalable and optimized framework with extremely high performance. It provides an ensemble of decision trees to ultimately provide better predictions. It utilizes gradient boosting to minimize the loss function and iteratively updates the weights of weak models. XGBoost has proved in several classification tasks to be superior, other than machine learning algorithms, because of regularization and parallel processing.

Long Short-Term Memory (LSTM) Network: A variant of the RNN, LSTM is a network designed to learn long-range dependencies in sequential data, like text. Since LSTMs avoid the traditional problem with RNNs, this network also employs gates that control the flow of information in order to preserve long-term context. Input to the LSTM network is tokenized and padded sequences of words. It would then process these sequences one word at a time while keeping an internal memory, or state, of the words seen thus far, which would allow it to capture sequential patterns and context.

Recurrent Neural Network (RNN): A Recurrent Neural Network (RNN) is yet another deep learning model suited for sequence data. Unlike most models, RNNs process inputs in the order received and have some sort of memory; it updates its hidden state after each word in the sequence. This makes RNNs particularly well-suited for tasks like sentiment analysis, where the meaning of a word may depend on words that came before it. Although much less complex than the LSTMs, the RNNs are yet capable of learning the temporal dependency information within the data, but they suffer from the vanishing gradient problem when learning long sequences.

Bayesian Network: A Bayesian Network is a probabilistic graphical model that encodes the conditional dependencies among a set of variables by a directed acyclic graph. It is proposed in this paper to construct a simple Bayesian Network for the sentiment prediction process. The Bayesian Network computes the probability distribution over possible sentiment labels (positive or negative), given a review, considering various features and their dependencies. Compared to black box models like deep neural networks, this model can offer insights into the likelihood of a sentiment prediction and hence is more interpretable.

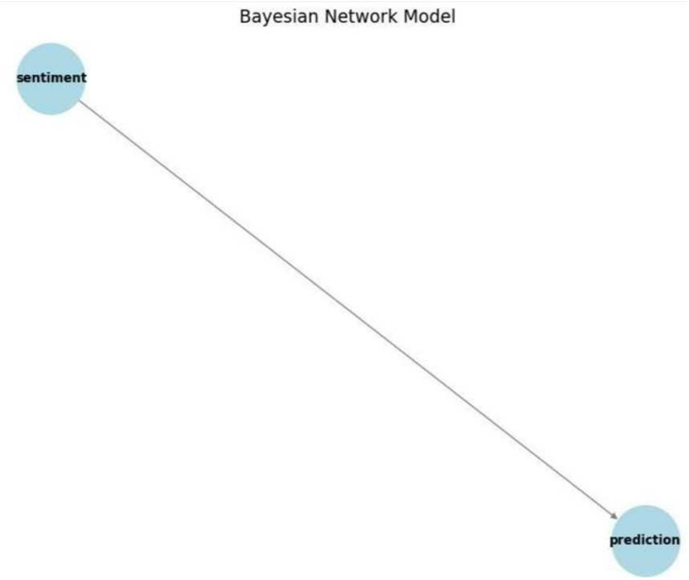
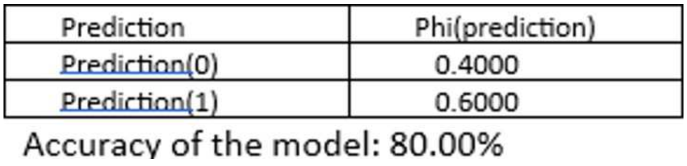


Fig. 4. Bayesian Network Model



Prediction	Phi(prediction)
Prediction(0)	0.4000
Prediction(1)	0.6000

Accuracy of the model: 80.00%

D. Train-test split

The dataset was split into training and testing sets, within which the performance of sentiment analysis models were ascertained. In this research, 60 percent has been used for training, and the rest, 40 percent, has been kept aside for testing purposes. This will split the data such that it trains the models on a big enough part of the data and uses the test set only as an independent set for validating how well the models generalize to unseen data. The splitting was done using train and test split from Scikit learn with the random seed to set for results reproducibility. Training was performed on the training set and the results were calculated in terms of performance metrics such as accuracy, precision, recall, F1 score, and ROC-AUC while making predictions on the test set.

E. Performance Evaluation

The trained models are evaluated using the test portion of the dataset, with performance measured through metrics like accuracy, precision, recall, and F1 score. Precision, for instance, represents the proportion of correctly predicted positive instances out of all instances predicted as positive, calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP).

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: Total number of relevant instances that were retrieved.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1- Score: The F1 score is a measure that combines precision and recall, providing a balance between the two by calculating their harmonic mean.

$$\text{F1 Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Accuracy: Accuracy measures how often the model correctly predicts the sentiment out of all the predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Model comparison using evaluation metrics

Models are compared based on key evaluation metrics. Those that include accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve), which will give a clearer performance in understanding. Accuracy sums up to give the general correctness of the predictions made, while precision measures the number of true positives identified out of all the positives reported by a model. This way, false positives are minimized. Recall determines all the true positives recognized by the model, therefore reducing false negatives. The F1-score balances precision with recall by providing just one score for the measure of performance on both measures. Finally, A better performance is represented by a higher value. All the above metrics can also be represented in visual form that may help in gaining further insights regarding model performance. Accuracy, precision, recall, and F1-score plots have a bar plot relative to strengths of each model. A model may have high precision while Naive Bayes has a high recall. A separate bar plot of AUC scores plots discrimination, with higher values closer to 1 being better. Thus, an analysis of trade-offs between metrics helps pick the best model for generalization.

RESULTS AND ANALYSIS

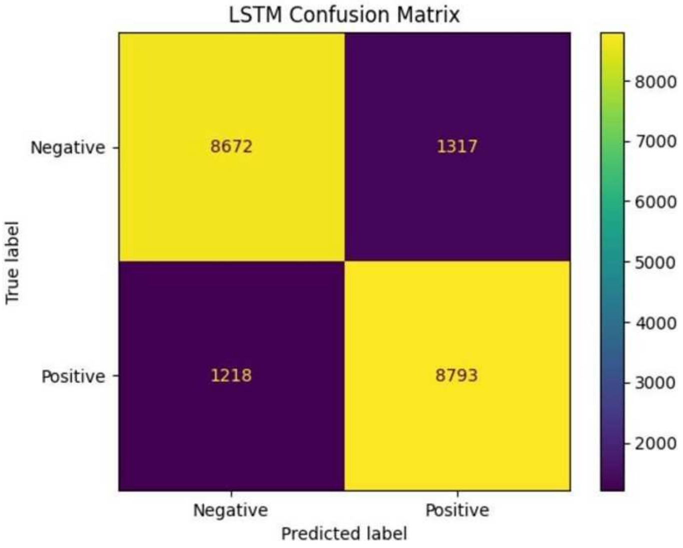


Fig. 5. LSTM Confusion Matrix

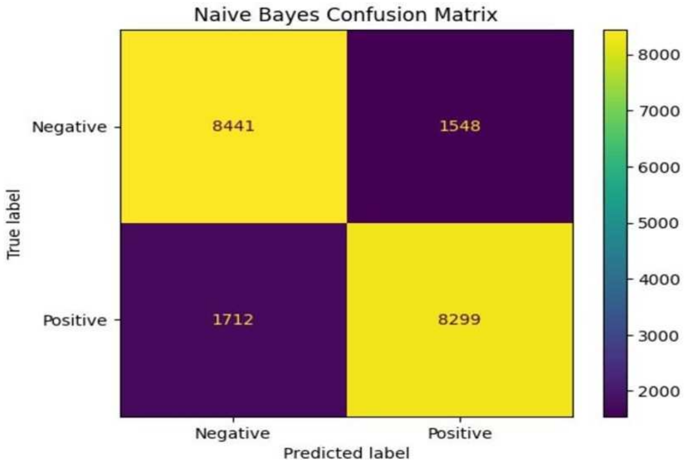


Fig. 6. Naive Bayes Confusion Matrix

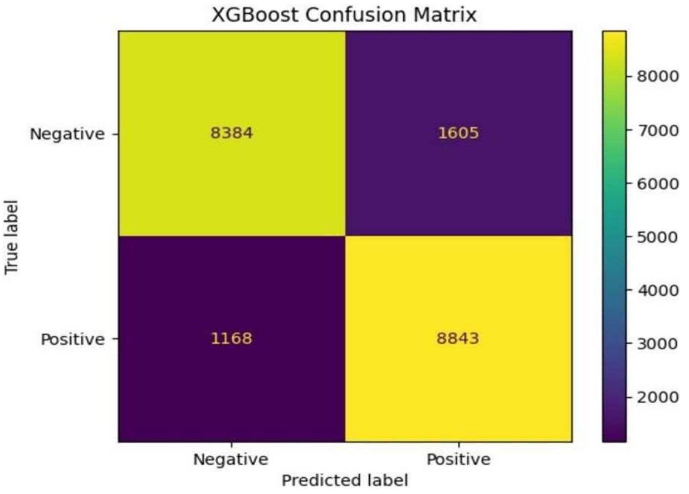


Fig. 7. XGBoost Confusion Matrix

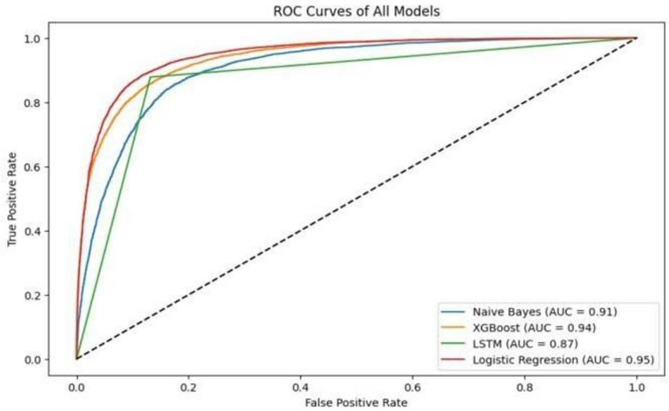


Fig. 8. ROC Curve for LSTM, Naive Bayes, XGboost and logistic regression Models

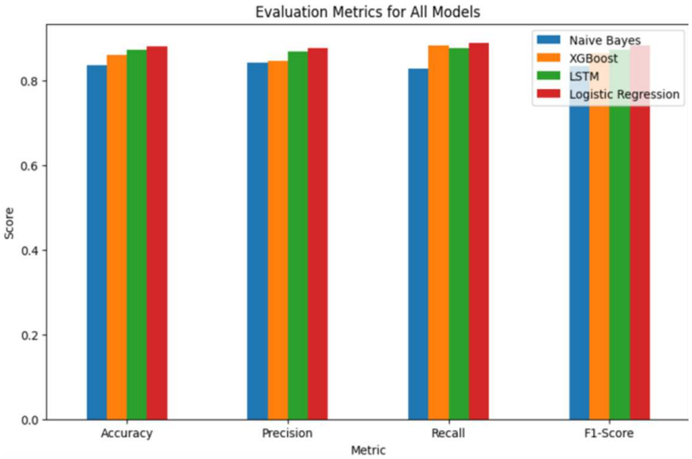
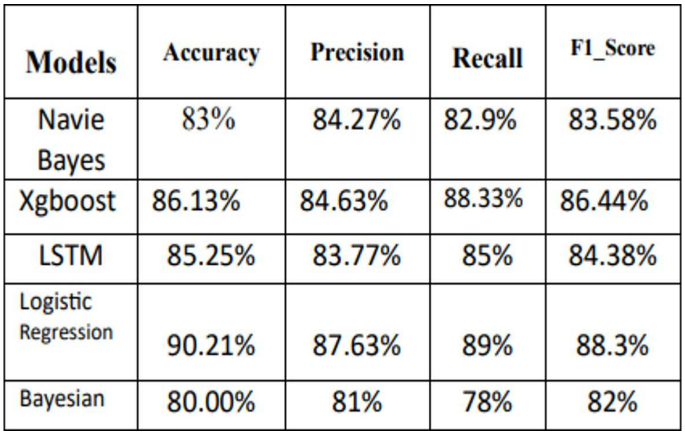


Fig. 9. Models Comparison of Evaluation metric



Models	Accuracy	Precision	Recall	F1_Score
Navie Bayes	83%	84.27%	82.9%	83.58%
Xgboost	86.13%	84.63%	88.33%	86.44%
LSTM	85.25%	83.77%	85%	84.38%
Logistic Regression	90.21%	87.63%	89%	88.3%
Bayesian	80.00%	81%	78%	82%

Fig. 10. Evaluation Metrics of Various Models

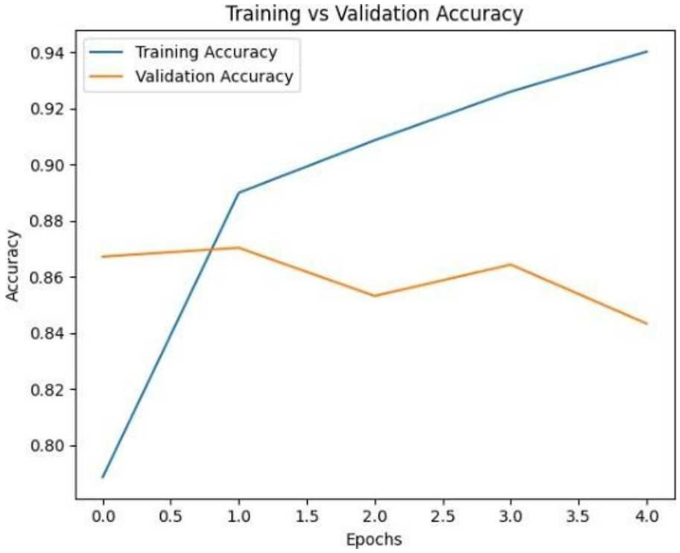


Fig. 11. Training vs. Validation Accuracy

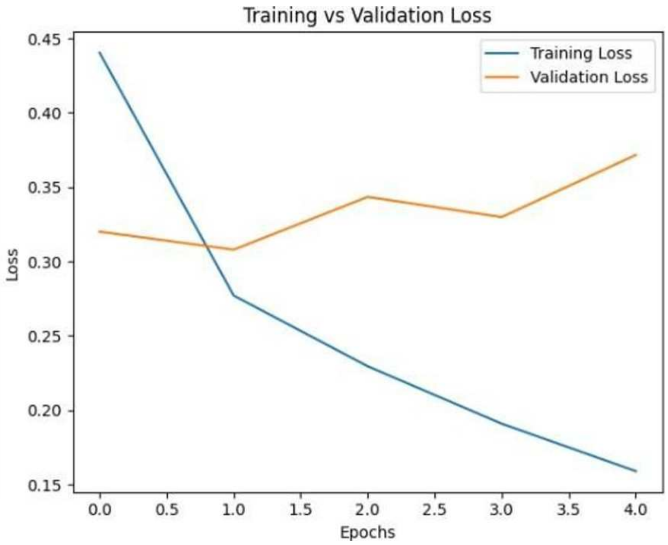


Fig. 12. Training vs. Validation Loss

CONCLUSION

It was shown that various ml and dl models work well on the sentiment analysis of IMDb movie reviews. Naive Bayes, for instance, took a Bag of Words approach at reasonable computational efficiency and interpretability. The model based on the gradient- boosting technique, XGBoost proved quite stable as far as its accuracy and precision are concerned. Among all deep learning models, the best performance was obtained by the LSTM network because it is particularly capable of capturing sequential patterns pertinent to the problem of natural language processing. Key findings: The classical models are computationally efficient but fail to capture the intricate nature of sequential data; in contrast, the deep learning model, most especially LSTM, provides better accuracy but at a higher level of computational complexity. Furthermore, inclusion of

Bayesian Network Model produces an element of probabilistic reasoning without any comparison to any other model. The hybrid frameworks can then integrate these models and be used in real-world applications in the near future with improved scalability and robustness.

References

[1] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, “A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis,” *Multimedia Tools and Applications*, vol. 78, pp. 26597–26613, 2019.

[2] A. Pandey, R. Yadav, A. Pathak, N. Shivani, B. Garg, and A. Pandey, “Sentiment analysis of imdb movie reviews,” in *2024 First International Conference on Software, Systems and Information Technology (SSIT-CON)*, pp. 1–6, 2024.

[3] J. Wang, H. Xie, F. L. Wang, L.-K. Lee, and M. Wei, “Jointly modeling intra- and inter-session dependencies with graph neural networks for session-based recommendations,” *Information Processing Management*, vol. 60, no. 2, p. 103209, 2023.

[4] B. A. B. Palomo, F. H. V. Velarde, F. J. Cantu-Ortiz, and H. G. Ceballos Cancino, “Sentiment analysis of imdb movie reviews using deep learning techniques,” in *Proceedings of Eighth International Congress on Information and Communication Technology* (X.-S. Yang, R. S. Sherratt, N. Dey, and A. Joshi, eds.), (Singapore), pp. 421–434, Springer Nature Singapore, 2024.

[5] N. S. L. S. Charitha, K. Yasaswi, V. Rakesh, M. Varun, M. Yeswanth, and J. S. Kiran, “Comparative study of algorithms for sentiment analysis on imdb movie reviews,” in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 824–828, 2023.

[6] A. Kumar, “Enhancing sentiment analysis accuracy in imdb movie reviews with deep learning models: A comparative study of rnn and cnn models,” in *2023 IEEE International Carnahan Conference on Security Technology (ICCST)*, pp. 1–9, 2023.

[7] M. Mahyarani, A. Adiwijaya, S. A. Faraby, and M. Dwifebri, “Implementation of sentiment analysis movie review based on imdb with naive bayes using information gain on feature selection,” in *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, pp. 99–103, 2021.

[8] S. M. Qaisar, “Sentiment analysis of imdb movie reviews using long short-term memory,” in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–4, 2020.

[9] K. Amulya, S. B. Swathi, P. Kamakshi, and Y. Bhavani, “Sentiment analysis on imdb movie reviews using machine learning and deep learning algorithms,” in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 814–819, 2022.

[10] S. Tripathi, R. Mehrotra, V. Bansal, and S. Upadhyay, “Analyzing sentiment using imdb dataset,” in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 30–33, 2020.

[11] P. Panda, R. Vatsa, D. Chime, and O. O. Awe, *Understanding Sentiment Analysis: Insights from Student Reviews of Universities and IMDb Movie Reviews*, pp. 341–372. Cham: Springer Nature Switzerland, 2025.

[12] S. Sabba, N. Chekired, H. Katab, N. Chekkai, and M. Chalbi, “Sentiment analysis for imdb reviews using deep learning classifier,” in *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, pp. 1–6, 2022.

[13] Q. A. Xu, V. Chang, and C. Jayne, “A systematic review of social media-based sentiment analysis: Emerging trends and challenges,” *Decision Analytics Journal*, vol. 3, p. 100073, 2022.

[14] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem, and T. Mahmood, “Sentiment analysis on imdb using lexicon and neural networks,” *SN Applied Sciences*, vol. 2, no. 2, p. 148, 2020.