Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis in Film Reviews

Mohamed Cherradi*, Anass El Haddadi

Data Science and Competetive Intelligence Team (DSCI), ENSAH, Abdelmalek Essaâdi University (UAE), 93000 Tetouan, Morocco

* Correspondence: Mohamed Cherradi (m.cherradi@uae.ac.ma)

Received: 05-08-2024 **Revised:** 07-10-2024 **Accepted:** 07-17-2024

Citation: M. Cherradi and A. El Haddadi, "Comparative analysis of machine learning algorithms for sentiment analysis in film reviews," *Acadlore Trans. Mach. Learn.*, vol. 3, no. 3, pp. 137–147, 2024. https://doi.org/10.56578/ataiml030301.

Abstract: Sentiment analysis, a crucial component of natural language processing (NLP), involves the classification of subjective information by extracting emotional content from textual data. This technique plays a significant role in the movie industry by analyzing public opinions about films. The present research addresses a gap in the literature by conducting a comparative analysis of various machine learning algorithms for sentiment analysis in film reviews, utilizing a dataset from Kaggle comprising 50,000 reviews. Classifiers such as Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Classification (LinearSVC), and Gradient Boosting were employed to categorize the reviews into positive and negative sentiments. The emphasis was placed on specifying and comparing these classifiers in the context of film review sentiment analysis, highlighting their respective advantages and disadvantages. The dataset underwent thorough preprocessing, including data cleaning and the application of stemming techniques to enhance processing efficiency. The performance of the classifiers was rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score. Among the classifiers, LinearSVC demonstrated the highest accuracy at 90.98%. This comprehensive evaluation not only identified the most effective classifier but also elucidated the contextual efficiencies of various algorithms. The findings indicate that LinearSVC excels at accurately classifying sentiments in film reviews, thereby offering new insights into public opinions on films. Furthermore, the extended comparison provides a step-by-step guide for selecting the most suitable classifier based on dataset characteristics and context, contributing valuable knowledge to the existing literature on the impact of different machine learning approaches on sentiment analysis outcomes in the movie industry.

Keywords: Artificial intelligence; Machine learning; Natural language processing; Sentiment analysis

1 Introduction

The world of movies has changed significantly, where films are no longer only for amusement but also participate in advertising and cultural discussions. Nowadays, a film's success is judged by its earnings and how it influences people's beliefs and buying habits [1, 2]. Although movie reviews are essential in shaping perception and helping viewers decide what to watch, current methods of analyzing sentiments often struggle to grasp the connection between audience feelings and the reception of a film [3, 4].

In order to overcome this restriction, this study employs machine learning methods to delve into and examine these complex dynamics. Classifiers, such as Logistic Regression, Naive Bayes, LinearSVC, and Gradient Boosting, were employed because of their capacity to manage and interpret extensive amounts of text data with enhanced complexity. This strategy allows for an exploration of emotions, uncovering patterns and trends that basic techniques might miss.

The major significance of this study is that it enhances the analysis and understanding of responses to movies, which can be useful for film industry practitioners as well as researchers. This study makes a unique contribution by comparing the machine learning algorithms to establish the best approach for sentiment analysis in movie reviews. This comparison also provides suggestions on algorithm selection depending on datasets and contexts.

This study's practical significance can be enhanced by providing better means to analyze and comprehend the emotional responses of movies, thereby benefiting professionals in the sector and academicians. The innovative contribution of this study lies in a comprehensive appraisal of these machine learning approaches, which identifies the

most effective model for sentiment analysis in film reviews and provides insights on how to choose the best algorithm among different datasets and environments. After choosing and preparing a dataset, machine learning algorithms were compared in this study. Then several metrics, such as accuracy, precision, recall, and F1-score, were used to evaluate the performance of these models. Finally, an informed choice of algorithm was made depending on those assessment metrics.

The remainder of this study is structured as follows: Section 2 presents a comprehensive review of related works, highlighting existing research in sentiment analysis of film reviews and machine learning methodologies employed in similar studies. Section 3 describes the research methodology in detail, encompassing the selection and preprocessing of the dataset as well as the application of machine learning algorithms for sentiment analysis. Section 4 presents the results of the study, along with in-depth discussions and an analysis of the findings. Finally, in Section 5, the conclusion is drawn, summarizing key insights gained from the research and offering potential avenues for future exploration and enhancement in the field of sentiment analysis of film reviews using machine learning approaches.

2 Related Works

In 2018, Muliono and Tanzil [5] proposed three approaches for text classification: k-Nearest Neighbor (kNN), Naive Bayes, and Support Vector Machine (SVM). They clearly demonstrated the usefulness of these algorithms in classifying datasets and the nuanced connections between (learned) algorithmic techniques and specific dataset attributes. In addition, they contributed to the idea that any classifier can perform well or poorly on any document collection, laying the stage for future research in text classification to be delegated to another method.

Subsequent studies, including those by several researchers [6–9], focused on the sentiment analysis of the movie reviews on the Internet Movie Database (IMDb), aiming to investigate multiple layers of sentiment in the reviews for evaluating the performance of movies, among other aspects. The application of machine learning techniques in these studies has facilitated the categorization of movies based on sentiment analysis conducted by critics, providing illustrative insights into the efficacy of these methods. In addition, it can be seen that the proposed models work fine on large sets of data, which aid in mapping the variety of responses that are built-in with emotion. These studies contribute to advancing sentiment analysis but have not explored the consequences of the various outcomes of using different machine learning models in sentiment classification.

In 2022, Alotaibi and Al-Rasheed [10] conducted a comparative analysis of sentiment analysis models using Twitter data, employing both lexicon-based and polarity multiplication approaches. This research underscored the importance of a model's ability to process large volumes of data with high precision and offered insights into effective strategies for analyzing social media datasets.

Built on these foundational works, this study comprehensively compares multiple machine learning classifiers, i.e., Logistic Regression, Naive Bayes, LinearSVC, and Gradient Boosting, specifically for sentiment analysis of movie reviews. Unlike previous studies, the proposed approach rigorously evaluates and compares these algorithms' performance, addressing the gap in comparative analysis and providing practical guidance for selecting the optimal model for sentiment analysis in the film industry. However, this comparative analysis not only advances the field by identifying the most effective algorithm for classifying sentiments in movie reviews but also contributes to a more nuanced understanding of sentiment analysis methodologies, enhancing both theoretical and practical applications.

3 Research Methodology

3.1 Proposed Methodological Framework for Sentiment Analysis in IMDb Reviews

Embarking on the intricate task of sentiment analysis within IMDb reviews, the proposed methodological framework underscores a fusion of meticulous data curation and cutting-edge computational paradigms. Illustrated in the accompanying schematic, the proposed methodology provides guidance through a rigorous exploration of the textual landscape, where raw feedback undergoes a transformative process of refinement. By employing advanced data preprocessing techniques, extraneous noise was systematically eliminated and the essence of sentiment embedded within the corpus was distilled. Taking advantage of state-of-the-art vectorization techniques, text-based descriptions were converted into vectorized feature spaces, amenable to analyses by computational tools and methods. Then this carefully picked dataset distribution was further split into specific sets for training and testing, respectively, beneficial for fine-tuning and testing of deep-learning models. Exploiting a vast array of the most up-to-date architectures of sentiment analysis, including those proposed by several researchers [11–15], with each methodically trained and tested, the proposed methodological excursion was finished with a comprehensive performance comparison to reveal how to achieve better sentiment analysis.

Figure 1 highlights a methodological process designed meticulously to obtain meaningful sentiments from the IMDb reviews. It is a chain of strictly defined procedures, each of which is aimed at improving the quality of raw textual data to produce actionable intelligence. The following illuminates these steps in detail:

• Data collection: Sentiment analysis was conducted using data from the online platform Kaggle. The dataset involves the use of movie reviews from IMDb and consists of fifty thousand reviews. This corpus was split into

two classes: 25,000 negative reviews and 25,000 positive ones. Each review is associated with two columns, with debatable tokens of 'review' and 'sentiment'.

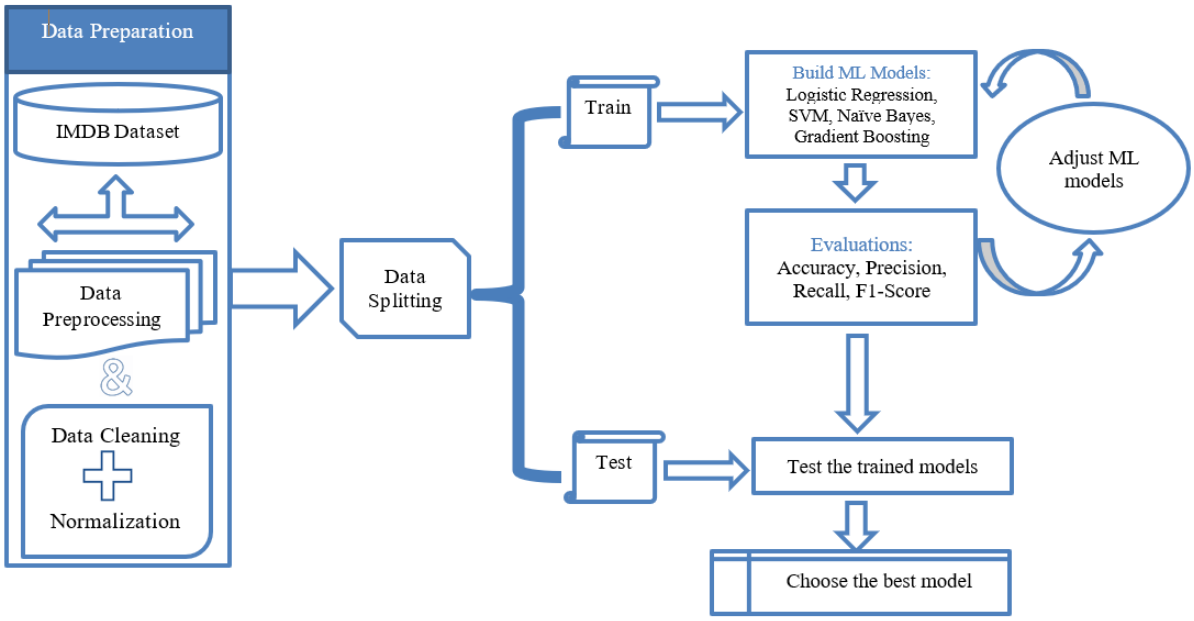

Figure 1. Proposed methodologies for film review sentiment analysis using machine learning

• Data cleaning and preprocessing: Data was preprocessed to convert it into lower case, with HTmachine learning tags, special characters, and punctuation stripped out. In the preprocessing, specific measures were taken when identifying the role of numerical digits. In many cases, numbers were removed to reduce noise and relevant numbers were specifically retained when they were deemed important, such as ratings or dates. Common stop words, Twitter handles, and hashtags were filtered out to focus on the core movie review content. Stemming normalized word variations, while tokenization segmented text for granular analysis. Vectorization relied on term frequency-inverse document frequency (TF-IDF) with n-grams to numerically represent text, considering term importance and capturing contextual information, thus enhancing the model's discriminative power.

• Data splitting: After preprocessing and transforming the data into a format suitable for analysis, the data was divided into three subsets: training, testing, and validation. In this study, approximately 70% of the data was allocated for training, 15% for testing, and 15% for validation. This ratio is commonly used in machine learning due to its balance between providing enough data for effective training and reserving a sufficient portion for unbiased evaluation and model fine-tuning. The training set was used to train the machine learning model, while the testing set was used to evaluate the model's performance on unseen data. The validation set together with the test set is yet another disjointed set, which helps adjust some of the parameters of the model to avoid overfitting. Nonetheless, it should be mentioned that different research situations may call for other ratios and the selection of the allocation can be modified according to the circumstances of a certain project, the accessibility of the information, and the level of sophistication of the models under consideration. To strengthen the increased operational reliability and versatility of the model, k-fold cross-validation was used with k = 5. The idea is to divide the data about the problem into five sets, also known as 'folds'. In each fold, one fold of data was left out, which is termed the validation set. The rest of the data, which is nearly 80 percent of the whole dataset, went into the training dataset. Each of the folds was only used once as the validation set, and the process was done five times. In general, cross validation gives a more accurate estimate of the generalization of the model than using only one dataset, which decreases variability, thereby better estimating the test data generalization of the model. This gentle way of splitting and cross-checking the data guarantees a fair assessment of the model's credibility, which gives better and more accurate machine learning results.

• Data modeling: This study considers systematic ways to perform sentiment analysis using many classification models, including Logistic Regression, Multinomial Naive Bayes, LinearSVC, and Gradient Boosting. All the models in data modeling frameworks have different strengths and characteristics that enhance the analytical process, making it more accurate in sentiment interpretation.

• Model fine-tuning: Improving the models is critical to making the sentiment analysis effort productive. To avoid overfitting and generalizability, several techniques were used to regularize model fit and output large coefficients with respect to the relative fitted functional form size. An early stopping approach was also used, which helps stop training

once the proposed model has attained the best score on a validation set, thereby avoiding overfitting by continuing to train too long. Moreover, cross-validation can assess how well the model generalizes over many different data splits. Optimizing various techniques, such as hyper-parameter tuning and feature representations, helps boost both the performance and accuracy of each classifier. By iteratively modifying the model parameters and configurations, its ability to generalize to previously unseen data was improved and the danger of overfitting was reduced. This methodical approach to model tuning ensures a comprehensive and dependable sentiment analysis pipeline capable of extracting subtle insights from a variety of textual data sources.

In the widespread context of IMDb reviews, the proposed creafted technique was used to conduct this pioneering study of sentiment analysis. This innovative approach to audience sentiment analysis challenges traditional boundaries because the proposed system raises the bar against conventional computational methods. Understanding the complex relationship between audiences and cinematic narratives can offer valuable cues for where sentiment analysis may be applied to film reviews.

3.2 Investigating Supervised Machine Learning Methods

This section focuses on the specific machine learning techniques used, especially the supervised ones used to identify feelings in the context of the IMDb reviews. The findings of this study support the use of four classification models to investigate sentiment analysis, i.e., Logistic Regression, SVM, Naive Bayes, and Gradient Boosting. These composite instruments were used to provide the maximum stratified results that subscribers wish to obtain. By employing these particular approaches, this study plans to obtain a more refined understanding of the audience disposition, thus enhancing the overall understanding of the methodological critique of cinema and digital communication in relation to the IMDb ratings.

3.2.1 Logistic regression

In the sphere of sentiment analysis, Logistic Regression plays a crucial role, as the method determines texts' sentiment as either positive or negative. It works under the principle that the likelihood of a piece of text being in either of the sentiment classes can be learned [16]. For example, in the process of classifying a movie review, the Logistic Regression determines to which category a review belongs, whether it is positive or negative. Therefore, the sigmoid function holds the centre stage in Logistic Regression, which transforms the scores of each text into a range of 0 to 1, a measure of the probability of positive sentiment. This function is mathematically defined as follows in Eq. (1):

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where, z is the sum of the weights of individual independent variables, such as the features of the given text, plus an additional bias term. This sum is then passed through a sigmoid function to make it a probability and then the model makes its decisions. Moreover, during the training process, Logistic Regression utilizes another approach known as the cost function in the estimation of the model's parameter. The cost function gives the difference between the probabilities that are expected and the sentiments that are actually assigned. Thus, the cost function is minimized, enabling Logistic Regression to make the correct predictions and categorize texts based on sentiment. Evidently, the cost function used in Logistic Regression is also referred to as the binary cross-entropy loss function or logarithmic loss function. Training example, on the other hand, is defined as a single piece of information as follows in Eq. (2):

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log \left(h_\theta \left(x^{(i)} \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - h_\theta \left(x^{(i)} \right) \right) \right] \tag{2}$$

where, $J(\theta)$ signifies the cost function, m indicates the number of training examples, $y^{(i)}$ represents the true label (0 or 1) of the i-th training example, $h_\theta \left(x^{(i)} \right)$ is the predicted probability that the i-th training example belongs to class 1, $x^{(i)}$ is the feature vector of the i-th training example, and θ is the parameter vector (weights) of the Logistic Regression model. When the model predicts a high probability for the incorrect class, this cost function penalizes the model more severely. The objective of the training process is to minimize this cost function by varying the parameter θ using methods such as gradient descent.

3.2.2 SVM

SVM is a powerful technique in sentiment analysis that can effectively classify text input into discrete sentiment classifications. Like Logistic Regression, SVM aims to delineate a decision boundary between different classes based on the features of the text [17]. At the heart of SVM lies the concept of maximizing the margin M between the decision boundary and the nearest data points, referred to as support vectors. Mathematically, the decision boundary is represented as in Eq. (3):

$$W^T \cdot x + b = 0 \tag{3}$$

where, W is the weight vector, and b is the bias term. The margin M is inversely proportional to the magnitude of $\|W\|$ subject to the constraint that all data points lie on the correct side of the decision boundary.

The optimization objective of SVM can be formulated as minimizing the following cost function (4):

$$\min_{w,b} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} \varepsilon_i \tag{4}$$

It is subject to the following constraints (5):

$$y_i \left(W^T \cdot x_i + b\right) \geq (1 - \varepsilon_i)\, avec\, \varepsilon_i \geq 0 \tag{5}$$

where, ε_i are slack variables that allow for some misclassification, C is a regularization parameter that balances the margin width and the classification error, x_i are the feature vectors, and y_i are the corresponding class labels. However, SVM employs kernel functions, such as linear, polynomial, or radial basis function (RBF) kernels, to map the input features into a higher-dimensional space where the data may be linearly separable. Thus, in summary, SVM optimizes the decision boundary by maximizing the margin between classes while penalizing misclassifications, making it a powerful tool for sentiment analysis tasks.

3.2.3 Naive Bayes

In the domain of sentiment analysis, Naive Bayes is a well-established algorithm used for classifying text data into different sentiment categories, such as positive or negative. It is a probabilistic classifier since it functions according to the ideas of the Bayes theorem [18]. The underlying premise of Naive Bayes is the conditional independence of features assigned a class label. Mathematically, this can be expressed as follows in Eq. (6):

$$P\left(x_1, x_2, \ldots, x_n \mid y\right) = P\left(x_1 \mid y\right) \times P\left(x_2 \mid y\right) \times \ldots \times P\left(x_n \mid y\right) \tag{6}$$

where, $x_1, x_2, \ldots, x_n$ are the features of the text, and y is the class label. Naive Bayes classification includes the calculation of the post-likely classifiers of every class, reflecting the features of the text and the selection of the most probable class. This is expressed as follows in Eq. (7):

$$P\left(y \mid x_1, x_2 \ldots x_n\right) = \frac{P(y) \times P\left(x_1 \mid y\right) \times P\left(x_2 \mid y\right) \times \ldots \times P\left(x_n \mid y\right)}{P\left(x_1, x_2 \ldots x_n\right)} \tag{7}$$

The prior probability $P(y)$ is the probability of sentiment y in the given dataset, whereas the likelihood $P\left(x_i \mid y\right)$ is the probability of the feature x_i given the class y.

Naive Bayes simplifies the probability calculation by assuming that features are conditionally independent, given the class label. Despite this 'naive' assumption, Naive Bayes often performs remarkably well in practice, especially for text classification tasks. Further, Naive Bayes computes the posterior probabilities of several sentiment classes using Bayes' theorem and chooses the class with the highest probability to be the predicted sentiment label for the provided text.

3.2.4 Gradient boosting

In the context of sentiment analysis, Gradient Boosting is a well-known ensemble learning algorithm used to train a base learner in order to categorize textual data based on given sentiments like positive, negative and so on. Unlike single models like Logistic Regression or Naive Bayes, Gradient Boosting combines multiple weak learners, typically decision trees, to create a strong predictive model [19]. At the core of Gradient Boosting is the concept of building a model step by step, adding one model boosting decision tree to the ensemble, with each added tree trained to reduce the errors made by the previous trees in the ensemble. Mathematically, this can be written as follows in Eq. (8):

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{8}$$

where, $F(x)$ is the final ensemble model, γ_m is the contribution of each tree $h_m(x)$, and M is the total number of trees in the ensemble.

Gradient Boosting's training process is based on finding an optimized loss function for the model, by building trees sequentially to fit the residuals between the target values and the previous set of predictions. This is done by using gradient descent optimization, in which the negative of the derivative of the loss function is used to make an update on the parameters of each tree. Thus, the Gradient Boosting model makes its final prediction by adding up all the individual trees' prediction weighted by their importance. Such a strategy enables Gradient Boosting to model dependencies in the data and deliver good predictive accuracy. Thus, Gradient Boosting forms a complex and powerful model of machine learning, composed of a set of relatively weak learners added sequentially, each of whom learns with the goal of minimizing the previous learner's errors. As such, the flow is cyclical and gives an extremely accurate sentiment classifier given large volumes of text data.

3.3 Assessment Metrics for Model Performance Evaluation

There are various possibilities to guide and optimize the training of a machine learning model. The model can also be evaluated within the proposed model analysis framework by applying a number of metrics [20, 21]. In this sentiment analysis task, efficiency was measured with the help of several parameters like accuracy, precision, recall, and F1-score. The important measures help assess the feasibility of the model and its efficiency in analyzing sentiment classifiers.

3.3.1 Accuracy

Accuracy defines the degree of correctness of a model and cannot be misled by any skewed classes like the F1 measure, which makes it the simplest yet most important forecast measure for classification models. Higher accuracy means that the fashioned model is close to perfect data prediction and lower values mean that it is dissimilar from the original observed values. However, accuracy may not give a total assessment of the performance of the model, especially when there are imbalances in the number of classes or when false positive and negative classes are equal with different ramifications. Therefore, although accuracy plays a significant role in evaluation, it is important that more indices should be incorporated to offer a full picture of the model's performance in practice. The accuracy formula is expressed as in Accuracy (%) = (Number of correct predictions) ÷ (Total number of predictions)×100 and in Eq. (9):

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{9}$$

where, TP denotes true positive, TN is true negative, FP denotes false positive, and FN is false negative.

3.3.2 Precision

Precision can be described as the percentage of actual positive labels that are correctly identified as such from the overall range of positive labels. They act as a key performance measure for determining the extent to which the model can correctly capture the positive elements within the dataset. A high precision score means that few of the confirmed negative cases are actually positive, thereby illustrating the ability of the predictive model to accurately identify positive cases with minimal misclassification. However, precision may not give a complete picture of the models' performance since the false negative rate is completely ignored. However, in scenarios where both false positives and negatives matter, precision should be combined with other metrics, such as recall and the F1-score, to give a clear indication of the model's performance in real-world use. The precision equation is formulated as follows in Eq. (10):

$$Precision = \frac{TP}{(TP + FP)} \tag{10}$$

3.3.3 Recall

Recall, or true positive rate, measures the capability of a model to predict positive in relation to all the positives in actuality. It is useful when predicting new data to ascertain the model's ability to identify all positive cases within the dataset, thus ensuring no positive instances are overlooked. Despite the richness of the recall in providing insights about the model's performance, it is crucial to consider other metrics like precision and the F1-score for a complementary understanding of the model's performance across different contexts and datasets. The recall equation is expressed as follows in Eq. (11):

$$Recall = \frac{TP}{(TP + FN)} \tag{11}$$

3.3.4 F1-score

The F1-score is another measure that integrates both precision and recall into one score, which gives a better view of the model. In general, the F1-score is calculated as the harmonic average of the measures of precision and recall, and varies from 0 to 1. This metric covers the capacity of the model to classify the positive instances with high precision and low numbers of false negative or positive results. This study has also shown that both precision and recall collectively reflect in the F1-score that is beneficial in achieving a balanced understanding of the model. This is because it assists in assessing the effectiveness of the specified model in real-life implementation. The F1-score equation is as follows in Eq. (12):

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{12}$$

Moreover, the confusion matrix is useful for the performance assessment of the sentiment analysis model because it shows all four types of errors possible in classification problems. Furthermore, the confusion matrix enables a

cross-tabulation of eleven classes, showing the number of true positives, true negatives, false positives, and false negatives. This matrix is especially important for the accuracy assessment of the methods in the definition of positive and negative points.

4 Results and Discussion

4.1 Experimental Design

The experiments were performed using Google Colaboratory (Colab), which is a freely available cloud-based tool with sufficiently powerful computational capabilities. In particular, the experiments were carried out on an HP EliteBook 8570p computer with an Intel Core i7-2600 CPU with 2. It features an Intel Core i7 2.2GHz 20GHz Intel Xeon processor, 16-Gradient Boosting RAM, and an Nvidia GeForce MX250 graphics card. The system runs on Windows 10 with a 64-bit architecture, accompanied by an Intel Core i7-3520M processor, which was selected to provide reasonable computing capacity and memory to process the large-scale dataset and execute sophisticated machine learning algorithms. In addition, the Python programming language was adopted for training the proposed machine learning models. The scikit-learn package was selected for machine learning based on its rich collection of tools and its simplicity. The experimental settings involved training the model with an Adam optimizer at a learning rate of 1e-4 for 175 epochs in total. This configuration was chosen in an attempt to increase the convergence speed of the layers and the accuracy of the model at the same time without overfitting. Therefore, the evaluation metrics of accuracy, precision, recall, and F1-score were selected as they offer an accurate insight into how well the models perform. Accuracy quantifies the general precision of a model; precision determines the true positive rate from the positive predictions; recall evaluates the essence of the model in terms of identifying all pertinent instances; and F1-score offers a blend of precision and recall scores. As a result, the use of these experimental settings and metrics is justified because they guarantee accurate and repeatable outcomes. The detailed performance results and their implications are presented in the subsequent sections, which contain descriptions about the utility of the presented model and the feasibility of the model in the sentiment analysis task.

4.2 Performance Evaluation

The findings depicted in Table 1 show the high accuracy of all four machine learning algorithms used when tested using the movie review set. Most notably, LinearSVC obtained the highest overall accuracy, equal to 90% of the total samples. With recognition accuracy of 89% for positive reviews and 91% for negative ones, the LinearSVC outperformed its counterparts and turned out to be a strong candidate for detecting true positive and negative cases. With an accuracy of 88% for the positive reviews and 89% for the negative ones, Logistic Regression closely followed LinearSVC with slightly lower accuracy. As for the parameter of recall, LinearSVC performed better than other algorithms, with a positive recall of 92% and a negative recall of 88%, indicating that LinearSVC also has a great ability to find a greater proportion of true positive and negative reviews. In terms of the recollection difference between the positive and negative reviews, Naive Bayes and Logistic Regression models performed mid-range and carried out the recollection at a comparable level. For F1-score, LinearSVC also performed significantly by keeping a balanced F1-score of 90% for positive and negative reviews. In addition, both Logistic Regression and Naive Bayes have comparable F1-scores. Although Gradient Boosting obtained low precision, it achieved a decent F1-score.

Table 1. Performance metrics of machine learning algorithms in sentiment analysis

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.881	0.885	0.871	0.876
Naive Bayes	0.871	0.880	0.890	0.885
LinearSVC	0.9098	0.912	0.894	0.896
Gradient Boosting	0.8064	0.828	0.768	0.797

Moreover, for every algorithm, the confusion matrices are shown in Figure 2, which compares the performance of the algorithms in the classification of positive and negative reviews. These matrices give a breakdown of the results of true positive, true negative, false positive, and false negative predictions. These confusion matrices are rather helpful for assessing the model's performance as they shed light on the kinds of classification errors. For instance, either high false positive or negative values suggest the need for the model's improvement in some aspect. Thus, the limitations of each algorithm were discussed based on the results from the confusion matrices. For instance, individual elements, such as the degree of combinations, the ability of generalization, and over- or under-fitting algorithms, may exhibit as a result. Overfitting is characterized by high accuracy on training data but significantly lower accuracy on unseen testing data; this is evident when a confusion matrix has a high variance. On the other hand, underfitting is indicated by poor performance within the training dataset and the test dataset in such a way that it portrays a model that does not capture the complexity of the underlying processes. It is also necessary to evaluate the generalization ability of

each model and assess the model's stability in different subsets of data samples. This also involves evaluating the model's performance in terms of the test or unseen data that is not part of the training set, which is essential for the model's usability. The aforementioned quantitative results are supported by the analysis of the confusion matrices provided in this study, which provides granular insight on the performance of each model for sentiment analysis tasks.

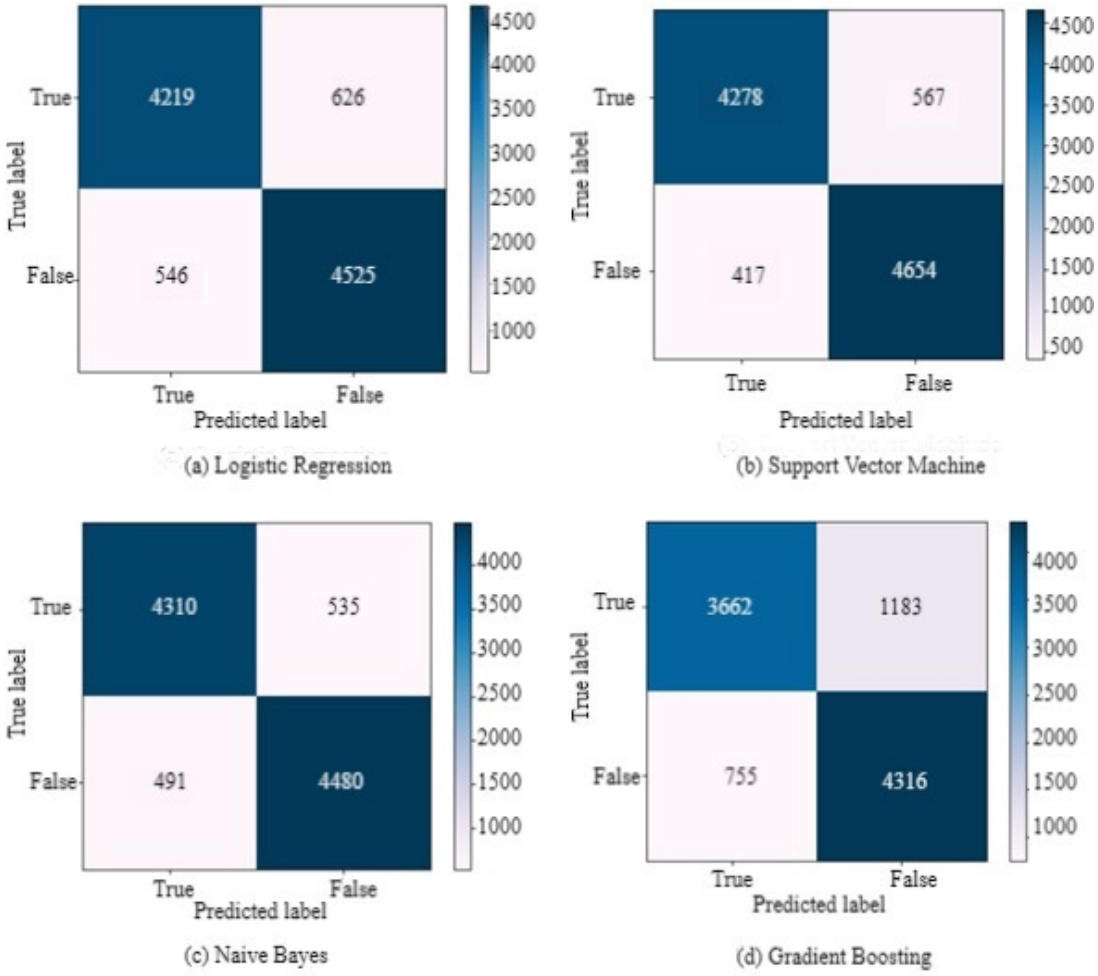

Figure 2. Confusion matrices for sentiment analysis algorithms

4.3 Enhancing Model Performance Through Hyper-Parameter Tuning

Hyper-parameter tuning serves as a significant advance in the advancement of machine learning models, as its goal is to adjust the enveloping settings of the primary algorithms as far as possible to enhance their performance on certain datasets [22, 23]. In this study, this methodology was used carefully and systematically across the proposed four algorithms to understand how it influences the effectiveness of all the given models. For Logistic Regression, the process began with the most significant hyper-tuning parameters like the regularization, the type or penalty, and the optimization solver. As for the Naive Bayes classifier, much attention was paid to the Alpha hyper-parameter. As for LinearSVC, the focus was placed on what kind of kernel to select from other parameters. Finally, as for Gradient Boosting, the optimization focused more on dividing the learning rate, the number of estimators, and a maximum depth. The efforts to fine-tune hyper-parameters provided spectacular results and revealed higher predictiveness for every model. Moreover, in order to strengthen the optimization process, a few extra techniques, such as Grid Search, were implemented, where all the pre-defined parameters were searched systematically to select the best parameters for each algorithm. This made the proposed models very specific by identifying all the necessary patterns from the dataset, which helped improve the performance of the models. However, it is best to visually display the improvements resulting from the tuning in order to best understand their impact. Therefore, Figure 3 depicts the 'before-and-after' accuracy, with a clearly sizable increase in the model's accuracy.

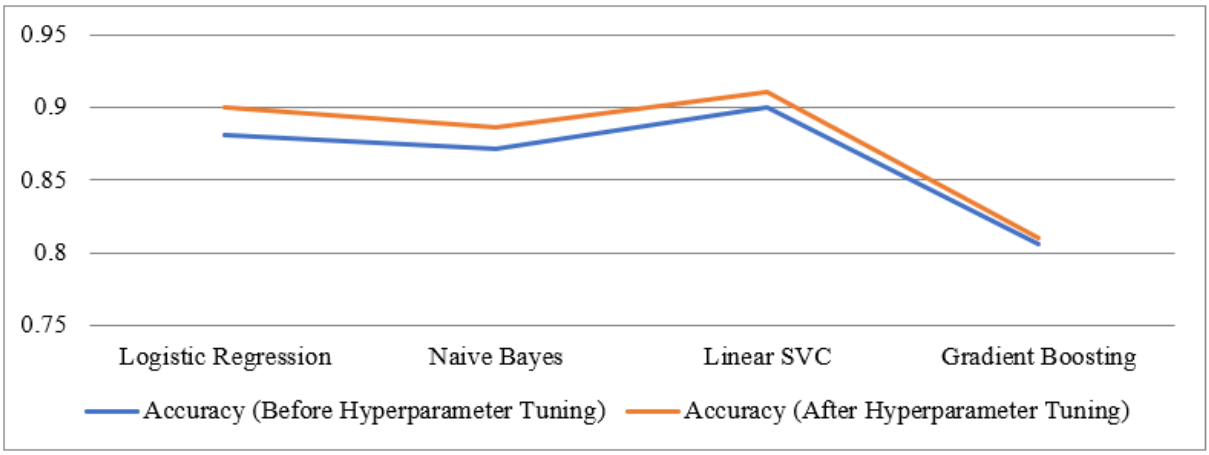

Figure 3. Accuracy improvement through hyper-parameter tuning

4.4 Discussion

This study is concerned with an exploration of the sentiment analysis field and trying to come up with a sound method that would allow classifying movie reviews as positive or negative. When applying the concepts of machine learning, different algorithms were initially experimented with and the hyper-parameters were adjusted. This study further extends prior research on sentiment analysis by implementing novel techniques to obtain relatively higher accuracy in sentiment classification problems. Therefore, by comparing them with related works in the area, the proposed approach showed remarkable improvements in the sentiment classification results. With the help of the most advanced methods of data preparation, feature set construction, and model selection, it is possible to reach high accuracy. Adoption of Logistic Regression in conjunction with LinearSVC and Gradient Boosting achieved an accuracy of 91%, which in turn enhanced the model's accuracy beyond the values obtained in earlier research. In order to make a clear distinction, Table 2 presents the studies of different authors, their methods used, and the accuracy levels achieved. The comparative analysis showed the efficient performance of this approach adopted in sentiment analysis, indicating that this approach can be applied practically in the near future to analyze the sentiments shared by audiences for movies.

Table 2. Comparison of sentiment analysis approaches

Algorithm	Accuracy	F1-Score
[24]	0.86	0.884
[25]	0.79	0.773
[26]	0.89	0.876
This study	0.91	0.897

5 Conclusion and Future Perspectives

In conclusion, this study has shown that machine learning techniques are highly effective for sentiment analysis in movie reviews. A significant aspect of this study is the thorough comparison of various machine learning algorithms, which helps pinpoint the best one for accurately classifying sentiments. This marks a significant advancement over existing techniques. Through meticulous development and fine-tuning, considerable strides have been made in distinguishing between positive and negative sentiments. The results of this study emphasize the potential of the proposed method to provide valuable insights into audience perceptions of movies. Furthermore, a new framework was introduced to evaluate sentiment analysis models, which combines quantitative metrics with an in-depth analysis of confusion matrices. This comprehensive evaluation enhances the understanding of model performance and offers a more solid assessment of its effectiveness.

Further, sentiment analysis models may be significantly enhanced in the future by various research endeavors. Using ensemble techniques, such as bagging and boosting, which combine the advantages of many algorithms to increase model resilience, is one possible avenue. Furthermore, investigating sophisticated deep learning architectures like transformers and recurrent neural networks (RNNs) may provide fresh perspectives on how to identify minute sentiment patterns in textual data. Context-aware characteristics and domain-specific knowledge may also be included to produce predictions that are more pertinent and accurate. Transfer learning techniques, where pre-trained models

are fine-tuned with movie review datasets, could speed up training and boost performance, especially when annotated data is limited. Finally, integrating multimodal data, such as combining text reviews with visual elements like movie posters or trailers, could provide a richer and more comprehensive understanding of audience sentiments. By addressing these future research directions, the field of sentiment analysis can be continuously advanced, contributing to its practical applications in various domains, including movie recommendation systems, market research, and social media analytics.

Author Contributions

Conceptualization, M.C., and A.E.; methodology, M.C., and A.E.; software, M.C.; validation, M.C. and A.E.; formal analysis, M.C.; investigation, M.C.; resources, M.C.; data curation, M.C.; writing—original draft preparation, M.C.; writing—review and editing, M.C.; visualization, M.C.; supervision, M.C. and A.E.; project administration, M.C.

Data Availability

The data for this study was sourced from a dataset obtained from the online platform Kaggle (https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset). The dataset used in this study is publicly available and can be accessed via the provided link.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] S. Chirgaiya, D. Sukheja, N. Shrivastava, and R. Rawat, "Analysis of sentiment based movie reviews using machine learning techniques," *J. Intell. Fuzzy Syst.*, vol. 41, no. 5, pp. 1–8, 2021. https://doi.org/10.3233/JIFS-189866

[2] P. V. Lakshmi, N. S. Kumar, G. Rao, and S. Medaka, "A deep learning approach for sentiment analysis of movie reviews," *Test Eng. Manag.*, vol. 83, 2020.

[3] A. Sajeevan and K. S. Lakshmi, "An enhanced approach for movie review analysis using deep learning techniques," in *International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2019, pp. 1788–1794. https://doi.org/10.1109/ICCES45898.2019.9002043

[4] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning techniques," *Int. J. Eng. Technol.*, vol. 7, no. 6, pp. 2038–2044, 2016.

[5] Y. Muliono and F. Tanzil, "A comparison of text classification methods k-NN, Naïve Bayes, and support vector machine for news classification," *Informatika IT*, vol. 3, no. 2, pp. 157–160, 2018. https://doi.org/10.30591/jpit.v3i2.828

[6] K. Amulya, S. B. Swathi, P. Kamakshi, and Y. Bhavani, "Sentiment analysis on IMDb movie reviews using machine learning and deep learning algorithms," in *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2022, pp. 814–819. https://doi.org/10.1109/ICSSIT53264.2022.9716550

[7] M. R. Haque, S. A. Lima, and S. Z. Mishu, "Performance analysis of different neural networks for sentiment analysis on IMDb movie reviews," in *International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, Rajshahi, Bangladesh, 2019, pp. 161–164. https://doi.org/10.1109/ICECTE48615.2019.9303573

[8] S. M. Qaisar, "Sentiment analysis of IMDb movie reviews using long short-term memory," in *International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2020, pp. 1–4. https://doi.org/10.1109/ICCIS49240.2020.9257657

[9] R. Talibzade, "Sentiment analysis of IMDb movie reviews using traditional machine learning techniques and transformers," in *Computer Science and Data Analytics*, 2023. https://doi.org/10.13140/RG.2.2.29464.16644

[10] S. Alotaibi and A. Al-Rasheed, "A review and comparative analysis of sentiment analysis techniques," *Informatica*, vol. 46, no. 6, pp. 33–44, 2022. https://doi.org/10.31449/inf.v46i6.3991

[11] Z. Kastrati, F. Dalipi, A. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students: A systematic mapping study," *Appl. Sci.*, vol. 11, no. 9, p. 3986, 2021. https://doi.org/10.3390/app11093986

[12] E. Kauffman, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining," *Sustain.*, vol. 11, no. 15, p. 4235, 2019. https://doi.org/10.3390/su11154235

[13] I. Lasri, A. Riadsolh, and M. ElBelkacemi, "Self-attention-based Bi-LSTM model for sentiment analysis on Tweets about distance learning in higher education," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 12, pp. 119–141, 2023. https://doi.org/10.3991/ijet.v18i12.38071

[14] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time Twitter sentiment analysis for Moroccan universities using machine learning and big data technologies," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 05, pp. 42–61, 2023. https://doi.org/10.3991/ijet.v18i05.35959

[15] E. S. Pereira, "Combining machine learning, lexical information and fuzzy technique to sentiment analysis for Brazilian Portuguese news," *Nat. Lang. Eng.*, vol. 1, no. 1, pp. 1–11, 2017.

[16] P. Reddy, D. Sri, C. Reddy, and S. Shaik, "Sentimental analysis using logistic regression," *Int. J. Eng. Res. Appl.*, vol. 11, no. 7, pp. 36–40, 2021.

[17] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment analysis using SVM: A systematic literature review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, 2018. https://doi.org/10.14569/IJACSA.2018.090226

[18] P. Chakriswaran, D. R. Vincent, K. Srinivasan, V. Sharma, C. Y. Chang, and D. G. Reina, "Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues," *Appl. Sci.*, vol. 9, no. 24, p. 5462, 2019. https://doi.org/10.3390/app9245462

[19] N. Subramani and D. Paulraj, "A gradient boosted decision tree based sentiment classification of Twitter data," *Int. J. Wavelets, Multiresol. Inf. Process.*, vol. 18, no. 4, p. 2050027, 2020. https://doi.org/10.1142/S0219691320500277

[20] N. Hicham, S. Karim, and N. Habbat, "Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 4, pp. 4504–4515, 2023. https://doi.org/10.11591/ijece.v13i4.pp4504-4515

[21] N. Panda, A. Jena, and V. R. Bendi, "Performance metrics assessment in sentimental analysis over machine learning approaches," in *IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2023, pp. 1–6. https://doi.org/10.1109/InC457730.2023.10263026

[22] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis," *Informatics*, vol. 8, no. 4, p. 79, 2021. https://doi.org/10.3390/informatics8040079

[23] E. Martínez-Cámara, N. Barroso, A. Moya, J. Fernández, E. Romero, and F. Herrera, "Deep learning hyper-parameter tuning for sentiment analysis in Twitter based on evolutionary algorithms," in *Federated Conference on Computer Science and Information Systems*, Leipzig, Germany, 2019, pp. 255–264. https://doi.org/10.15439/2019F183

[24] I. Steinke, J. Wier, L. Simon, and R. Seetan, "Sentiment analysis of online movie reviews using machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022. https://doi.org/10.14569/IJACSA.2022.0130973

[25] S. Samsir, K. Kusmanto, A. Dalimunthe, R. Aditiya, and R. Watrianthos, "Implementation Naïve Bayes classification for sentiment analysis on internet movie database," *Building of Inform. Tech. Sci.*, vol. 4, no. 1, pp. 1–6, 2022. https://doi.org/10.47065/bits.v4i1.1468

[26] K. Kaushik and M. Parmar, "IMDb movie data classification using voting classifier for sentiment analysis," *Int. J. Comput. Sci. Eng.*, vol. 10, no. 1, pp. 18–23, 2022. https://doi.org/10.26438/ijcse/v10i1.1823