SENTIMENT ANALYSIS OF MOVIE REVIEWS USING A STACK-BASED ENSEMBLE METHOD

Supported by tetfund

***O. O. SOLANKE, O. M. AZEEZ, K-K. A. ABDULLAH, S. O. HASSAN, F. E. AYO and M. A. USMAN**
Department of Computer Sciences, Olabisi Onabanjo University, Ago-Iwoye
*corresponding author: solanke.olakunle@oouagoiwoye.edu.ng

Received: February 14, 2025, Accepted: April 28, 2025

Abstract: Movie reviews are public reviews about films distributed via social media platforms, incorporating subjective impressions of viewers regarding their ratings. Class imbalance have created main problems to classifiers that are used in developing predictive models which tend to have high false positive rates as a result of the presence of a majority class found in the ratings dataset. This work, therefore, developed a classification model for an automated sentiment analysis of movie reviews, using stacked-based ensemble model of three machine learning classifiers (Decision Tree DT, Naive Bayes NB, and Logistic Regression LR). Two of the algorithms were used as 'base' and the third learner algorithm was used as 'meta' in rotating batches. Among the models developed using the isolated Machine learning algorithms (NB, LR and DT), LR performed best with total correct classification of 8840 and with an accuracy of 88.4%. Also, among the models developed using the stack-ensemble of Machine Learning algorithms, DT/NB as base learners and LR as the meta-learner, performed best with a total correct classification of 9408 and an accuracy of 94%. The study further emphasized the importance of adopting a stack ensemble of ML algorithms over isolated algorithms that usually have limiting capabilities, following which this adoption lowers the possible false positive rate likewise the rate of false negative rate of movie review sentiments. The exploration of other advanced sentiment analysis techniques is also recommended in this study.

Key words: Machine learning algorithms, Movie reviews, Sentiment analysis, Predictive models, Stacking ensemble, Meta-classifiers.

Introduction

The increase in the use of social media like blogs and social networks over the past decade, has led to a growing interest in sentiment analysis. With Web (2.0) revolution and the amount of user generated opinion data available online, personal views and opinions are no longer limited to newspaper writers or personalized opinion polls. Instead, almost anyone can express their opinions through social media. The abundance, availability, and accessibility of these opinions have given rise to automated applications that use sentiment analysis (opinion mining) as a key factor for stock market prediction, product reviews, service reviews, public opinion polls, and more. Sentiment analysis aims to uncover the differences between feelings such as happiness, sadness, grief experiences such as hatred or anger and emotions like love, as well as opinions expressed in reviews, online posts, and social media activity such as Facebook, Twitter, Linkedln, Instagram (Mao, et al.,2024; Almi and Simatupang, 2021; Gan, 2023).

Consequently, using opinion mining, one can determine the sentiment of movie reviews. Sentiment analysis can be complicated with the uses of misspellings, slang words, repeated character, abbreviations, local languages, and newly created emojis. The process of data mining has markedly diminished the difficulties connected with false positive and false negative evaluation. Discovery of knowledge within text database with the application of machine learning algorithms facilitate the identification of latent patterns to repository relationships among a multitude numbers of variables and predict sentiment outcomes based on historical opinions (Baid et al. 2017; Bing, 2020; Hamid and Abdulazeez, 2024).

Related works reveal that the problem of class imbalance is part of the main problems faced in the formulation and development of structured and effective predictive models (Waggle, et. Al., 2023). Class imbalance poses significant challenges for classifiers used to develop predictive models often leading to high false positive rate due to the dominance of the majority class in the collected dataset. However, a highly effective approach to overcoming class imbalance is to implement an ensemble model that combines two or more classifiers instead of depending on just a single classifier. To tackle these challenges, this research investigated the application of Stacking Ensemble models for classification of sentiment in movie reviews (Mao, et. Al., 2024).

Therefore, in order to increase the precision, accuracy and dependability of the sentiment analysis system, this study applied stack-based ensemble classifiers for the analysis of the sentiments of movie reviews of 50k IMDB on micro-blog (kaggle). The study developed suitable technique for preprocessing the collected dataset based on TF-IDF features extraction preparatory to analysis and compared the result of the research experiments with related results in the open literature.

Basarslan, et al. (2022), investigates the use of machine learning techniques including K-Nearest Neighbor, Random Forest, and Nave Bayes to scrutinize movie reviews and tweet sentiments. Feature extraction is performed using the StringToWordVector filter and Attribute Selection filter in WEKA, to prepare the data for classification.

The paper highlights the challenges encountered in sentiment analysis, encompassing slang and emotive memes, and presents the findings of an examination conducted on IMDB 2000 user-confirmed movie sentiment. Naive Bayes have the most accuracy at 81.4%. The study also reveals that, with an accuracy of 81.45%, the Naïve Bayes classifier works well for tweets, indicating the possibility for additional investigation into other algorithms to raise the precision of news analysis apps.

Adeyemi, et al. (2019), create a stack assemble model with datasets from the UCI machine learning repository online to classify patients with recurrent breast cancer. The preprocessing of their dataset was carried out methodically by using suitable bin intervals to transform all of the numerical features to nominal values. Three supervised learning algorithms was adopted in formulation of models used in the study, the Naive Bayes, Decision Tree and Support Vector Machine, application of two machine learning algorithms was formulated as base classifiers while meta-classifier was utilized by the third algorithm.

The simulation was done by using 10-way cross validation method for stack ensemble models on WEKA simulation environment, and the evaluation was carried out based on performance comparison and certain validation techniques. The models constructed using DT as the meta and NB/SVM as the base performed significantly better than other stack ensemble models, which used SVM and NB as the meta-classifier and performed equally with no differences. This led to the conclusion in their work that using DT as a composite classifier showed a more proficient capability than SVM and NB classifiers in classifying recurrence of breast cancer. However, forthcoming work in this filed could be substantially enhanced by adopting more larger datasets, incorporating additional features, exploring sophisticated machine learning approaches, and emphasizing on practical applications as well as interpretability.

Horsa and Tune (2023), argues, to close the gaps in earlier studies on this subject, create an aspect-based sentiment analysis (ABSA) model for Afaan Oromoo movie reviews using supervised machine learning approaches. The dataset utilized consisted of 2,800 manually annotated movie reviews of four different movies by Afaan Oromoo published between the year 2019 and 2021.

Machine learning algorithms like Random Forest, Naive Bayes, Support Vector Machine, and Logistic Regression are employed in the study, along with feature extraction methods like TF-IDF and Bag of Words (BoW). The study found that while MNB and SVM classifiers were very accurate, the Random Forest classifier outperformed the others, with an accuracy of 88%.

Kaushik and Parmar (2021) affirms that remarks from viewers of a film are known as movie reviews. These evaluations determine whether or not the film is worthy of seeing. In addition to providing information about the film's positive and negative qualities. Many reviews that people write and publish online can be helpful to other people. Considering how crucial this kind of information is for making decisions, a lot of individuals use the Internet.

Materials and Methods

This work employ sentiment analysis in the ensemble model. As shown in Figure 1, after the sentiment class labeled data was passed through preprocessing, a distinct text representation was produced using the word frequency based TF-IDF. Then, Level1: Machine Learning (DT, LR, and NB), were developed and trained so their predictions can be used as an input to develop the Level2: Stack Ensemble Learning and the results of these models were obtained.

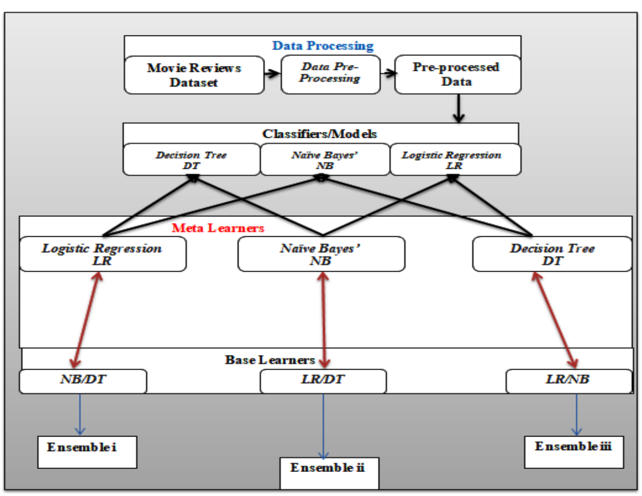


Figure 1: The Research Model

The Stack-Based Ensemble Model

A. Dataset:

The 50k IMDB movie reviews that were acquired from the Kaggle corpus served as the dataset for this study. The dataset was in Microsoft Excel format (.csv) which was an exact requirement of simulation environment used in this study. It has fifty thousand fiercely critical reviews, each with a strong sentiment. Reviews and sentiment are the two attributes that make up this bidirectional dataset. Reviews are listed in the review column, while the attitude column indicates whether or not the review is positive and negative. For each class, there are 25,000 records accessible. It is a balanced dataset with no null values.

B. Pre-Processing:

The unstructured movie reviews underwent preprocessing such as; Tokenizing, Normalizing, Vectorizing etc. to convert the contents of the review into word sets seen within each review, removal of stop words was also done from the extracted review contents. The retrieved contents were subjected to the stemming process in order to convert words such as "families" and "famili" into their base words. In order to execute the stemming of the terms retrieved from the reviews, Porter's Algorithms were applied to each document's words and was reduced to its root words, after which the frequency of occurrence of each word is taken into consideration. Figure 2 shows the experimental sample view of the natural language processing.

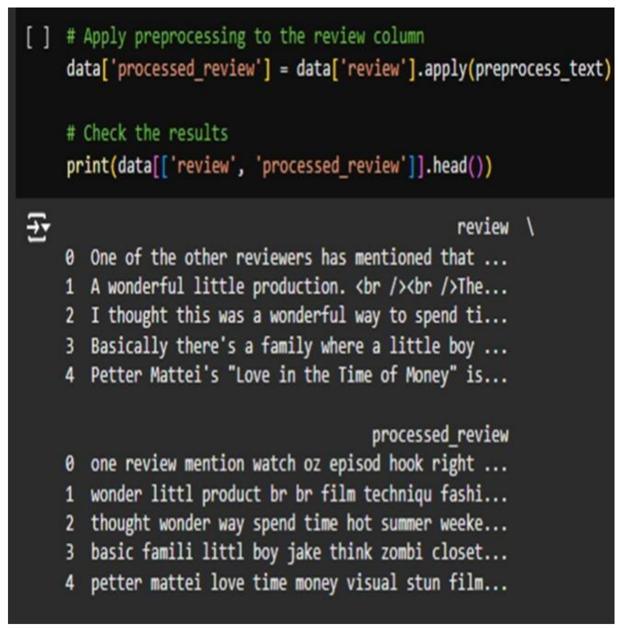


Figure 2: Sample of pre-processed data (Processed_review)

Classifier models

The Logistic Regression, Decision Tree, and Naïve Bayes constitute the stack ensemble model developed for this research. Two distinct machine learning algorithms were implemented as base learners within the framework of the study, while the third algorithm functioned as the Meta-learner, integrating them as input to construct the stack-ensemble model, which proved to be pivotal for the classification of sentiment in movie reviews. The Meta-learner was employed to determine how the predictions from the base learners were combined to achieve maximal classification accuracy. A detailed elucidation of the machine learning algorithms used in this study is provided, along with a description of how the algorithm, when functioning as a meta-learner, handles the two input base learners.

i. ***Decision Tree (DT):*** Decision Tree is used for its capacity for interpretability in handling non-linear relationships within the dataset. The decision trees methodology that was used to create the decision tree using the divide and conquer method is as follows, given Xij of j number of cases and i input features. Equation 2 shows the split ratio, which is used to determine which of the selected attribute splits is most effective in splitting the dataset after attribute selections by equation 1 gain ratio is determined by the division of equation (1) by equation (2), which are referred to as the information gain and the split criteria, respectively, in the two equations that the decision tree uses.

$$IG(X_i) = H(X_i) - \sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} H(X_i) \quad (1)$$

Where:

$$H(X_i) = -\sum_{t\epsilon T} \frac{|t,X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t,X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \quad (2)$$

T is the values of a given attribute X_i.

ii. ***Naive Bayes (NB):*** Selected due to its elegance and efficacy in domain of text classification task. The Naive Bayes Classifier, based on Bayes' theorem, operates as a probabilistic model. It is recognized as a statistical classifier that provides effective learning algorithms and insights from previously evaluated data. Let C represent the target class collected for j records, with X_{ij} being a dataset sample predicted by base classifiers i along with their corresponding event classes. $H_k = \{H_1=\text{Positive}, H_2=\text{Negative}\}$ represents a hypothesis that X_{ij} is associated with class C. In the context of classifying sentiment in movie reviews, given the input variable values for the jth record, the Naive Bayes classification is determined by the following steps:

a. $P(H_k \mid X_{ij})$ – Posterior probability: This denotes the probability that the hypothesis H_k holds, given the bserved data sample X_{ij} for $1 \leq k \leq 2$.
b. $P(H_k)$ - Prior probability: This indicates the initial probability of the target class, where $1 \leq k \leq 2$;
c. $P(X_{ij})$ This represents the probability of observing each attribute of the sample data, which in this case are the prediction outputs of base classifiers i; and
d. $P(X_{ij} \mid H_k)$ This denotes the probability of observing the sample's attribute X_{ij}, given that the hypothesis H_k holds in the training data

Therefore, in accordance with Bayes' theorem, equation (3) defines the posteriori probability of a hypothesis, while the class label that has maximum likelihood in equation (4) determines the classification of sentiment analysis of movie reviews of a record.

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2 \quad (3)$$

$$Class = MAX[P(Positive|X_k), P(Negatice|X_k)] \quad (4)$$

iii. ***Logistic Regression (LR):*** To get the ideal coefficients that minimize the loss function, logistic regression employs optimization techniques such as Gradient Descent and Iteratively Reweighted Least Squares (IRLS). For jobs involving binary classification, supervised machine learning techniques like logistic regression are employed. It is not a regression algorithm, despite its name, but a classification algorithm. The probability of a binary result (1/0, Yes/No, True/False) depending on one or more predictor variables (features) is predicted using logistic regression. The likelihood that a given input, XXX, belongs to a specific class is modeled using logistic regression. This function maps any real-valued number into a range between 0 and 1 using the logistic (sigmoid) function.

The logistic function (sigmoid function) is defined as:

σ(z) = 1/(1+e^(-x)) (5)

Where,

σ(z) = output in range 0 and 1; x = input;

e = base of nature log.

Results and Discussion

Performance Evaluation Metrics for Model Validation

The predictive model's effectiveness can be evaluated using the True positive or negative cases that were documented from the confusion matrix. The following express the definition and presentation of the metrics:

a. True Positive rate (TPR) known as sensitivity or recall is the proportion of true positive or negative which cases are correctly misclassified (Figure 3).

TP(Positive) = TP/(TP+FN) (6a)

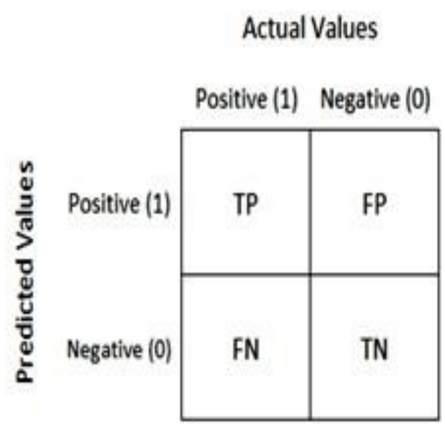


Figure 3: Model Performance Evaluation

TP(Negative) = TN/(TN+FP) (6b)

False Positive rate (FP) known as specificity 1/false alarms: This is the proportion of false positive or negative which are cases that are misclassified.

FP(Positive) = FP/(FP+TN) (7a)

FP(Negative) = FN/(FN+TP) (7b)

b. Precision is the proportion of predicted positive or negative cases that were correctly classified

Precision(Positive) = TP/(TP+FP) (8a)

Precision(Negative) = TN/(FN+TN) (8b)

c. Accuracy: This refers to as the proportion of total correct predictions

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \qquad (9)$$

Results and analysis

The simulation of the classification model developed for sentiment analysis was done using percentage split such that Eighty (80) percent of the dataset was used for model training, while twenty (20) percent was used for model validation. The trained dataset consists of 40000 reviews and the testing dataset consisted of 10000 reviews. Among the testing dataset, 5039 were positive reviews and 4961 were negative reviews. Based on the utilization of the testing dataset, the model validation results were presented.

Experiment I: Classification Performance and Confusion Matrix of Isolated Models

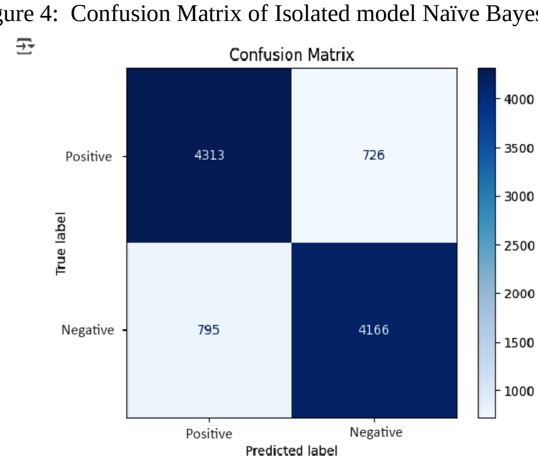
This section present the classification results of the classifiers adopted: Naive Bayes, Decision Trees (DT) and Logistic Regression (LR). Table 1, Table 2 and Table 3 as well as Figure 4, Figure 5 and Figure 6 detailed the comparison in term of evaluation performance and Confusion Matrices respectively for both isolated models versus stacked ensemble models.

a. Classification Report of Naive Bayes (NB)

Table 1: Classification Report of Naïve Bayes

	Precision	Recall	f1-score	Support
Negative	0.84	0.86	0.85	4961
Positive	0.85	0.84	0.85	5039
Accuracy			0.85	10000
Macro avg	0.85	0.85	0.85	10000
Weight avg	0.85	0.85	0.85	10000

Figure 4: Confusion Matrix of Isolated model Naïve Bayes (NB)



b. Classification Report of Decision Tree (DT)

Table 2: Decision Tree (DT) Classification Report

	Precision	Recall	f1-score	Support
Negative	0.72	0.72	0.72	4961
Positive	0.71	0.71	0.72	5039
Accuracy			0.72	10000
Macro avg	0.72	0.72	0.72	10000
Weight avg	0.72	0.72	0.72	10000

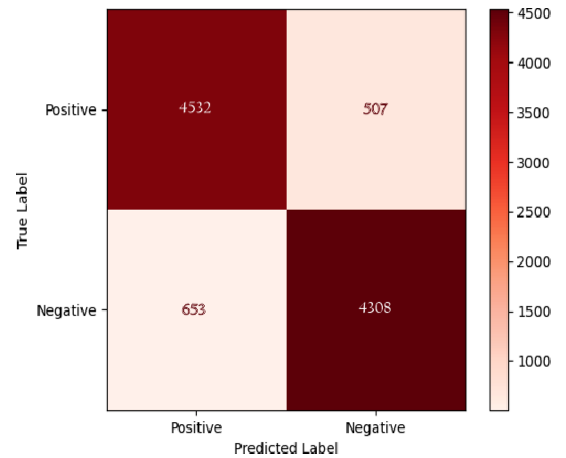


Figure 5: Confusion Matrix of Isolated model Logistic Regression (DT)

c. Classification Report of Logistic Regression (LR)

Table 3: Logistic Regression (LR) Classification Report

	Precision	Recall	f1-score	Support
Negative	0.87	0.89	0.88	4961
Positive	0.89	0.87	0.89	5039
Accuracy			0.88	10000
Macro avg	0.88	0.88	0.88	10000
Weighted avg	0.88	0.88	0.88	10000

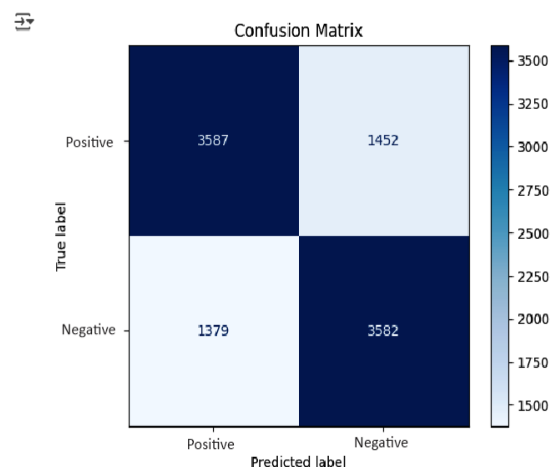


Figure 6: Confusion Matrix of Isolated model Logistic Regression (LR)

Experiment II: Classification Performance and Confusion Matrix of Stacking Ensemble Models

This section presents results of the Stacked Ensemble models developed for the study. As previously mentioned, the procedure was carried out such that the base learners used in the study employed two (2) classifiers, and the third classifier was designated as the meta-classifier. This process was done for all the classifiers identified for this study in batches which led to the development of 3 ensemble learning model for sentiment analysis classification.

The stack formulations of classifiers for this study are as follows:

i. Ensemble I: LR and NB as base model, DT as meta- model

(Table 4 and Figure 7)

ii. Ensemble II: DT and NB as base model, LR as meta-model (Table 5 and Figure 8)

iii. Ensemble 3: LR and DT as base model, NB as meta- model (Table 6 and Figure 9)

a. Ensemble 1: LR and NB as Base Models, DT as Meta- Model

Table 4: LR And NB as base Models, DT as Meta-Model

	Precision	Recall	f1-score	Support
Negative	0.84	0.84	0.88	4961
Positive	0.84	0.84	0.84	5039
Accuracy			0.84	10000
Macro avg	0.84	0.84	0.84	10000
Weight avg	0.84	0.84	0.84	10000

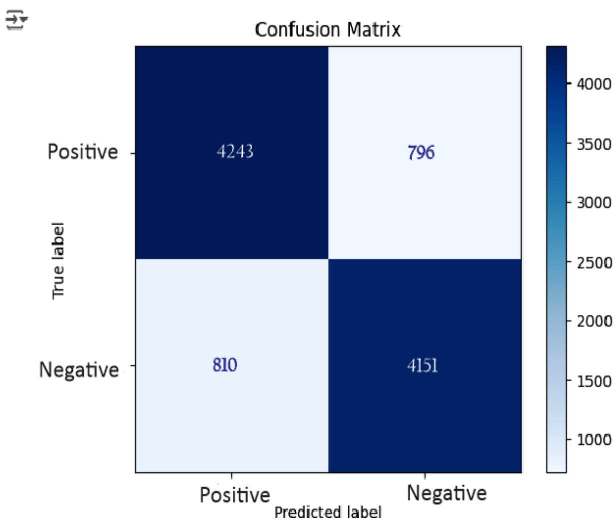


Figure 7: Confusion Matrix of Stacked Ensemble LR and NB as Base Models, DT as Meta-Mode

b. Ensemble 2: DT and NB as base Models, LR as Meta-Model

Table 5: DT And NB as base Models, LR as Meta-Model

	Precision	Recall	f1-score	Support
Negative	0.93	0.94	0.94	4961
Positive	0.94	0.94	0.94	5039
Accuracy			0.94	10000
Macro avg	0.94	0.94	0.94	10000
Weighted avg	0.94	0.94	0.94	10000

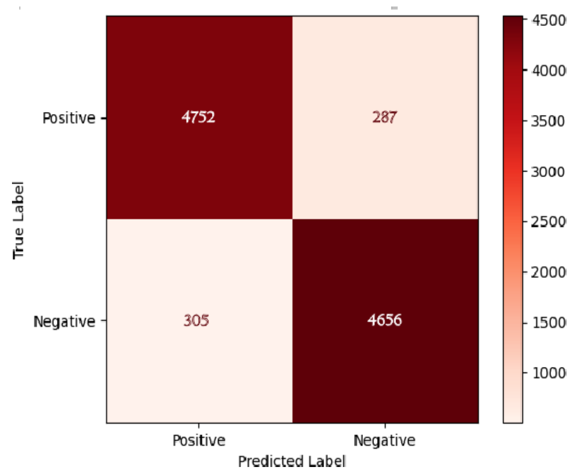


Figure 8: Confusion Matrix of Stacked Ensemble DT and NB as Base Models, LR as Meta-Mode

c. Ensemble 3: LR and DT as base Models, NB as Meta-Model

Table 6: LR and DT as base Models, NB as Meta-Model

	Precision	Recall	f1-score	Support
Negative	0.89	0.92	0.90	4961
Positive	0.92	0.89	0.91	5039
Accuracy			0.90	10000
Macro avg	0.90	0.90	0.90	10000
Weighted avg	0.90	0.90	0.90	10000

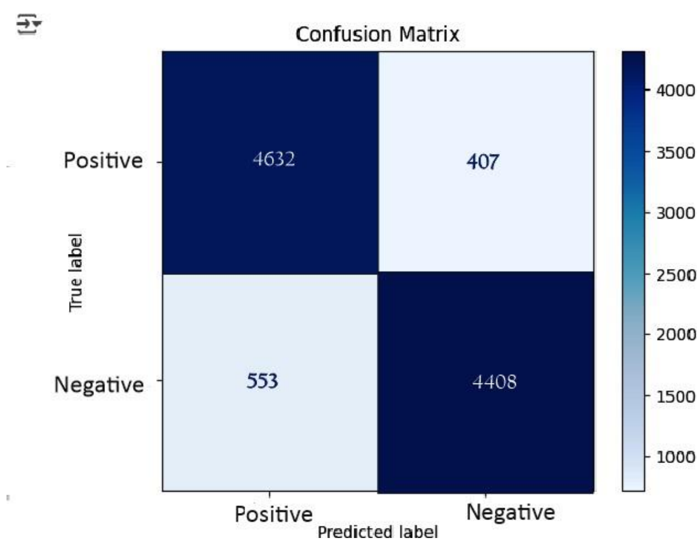


Figure 9: Confusion Matrix of Stacked Ensemble LR and DT as Base Models, NB as Meta-Model

Discussion of results

Figure 10 shows the empirical summary of experimental results of the developed models' classification performances carried out in this study. The table in the figure estimated and compared the results based on Accuracy, Precision, Recall, and specificity (f1-score) for both Isolated Models (Up) and Stacked Ensemble Models (Down) designed for this study. The highlighted results in Red color are the results of model that has high classification performance among the isolated models and Stacked Ensemble models respectively. The result justify the advantages of the Ensemble models in classification performances in sentiment analysis over the using of isolated classifiers which is by far flexible comparing to robustness performance of Ensemble models.

Isolated ML Algorithms	Correct Classification	Accuracy (%)	Recall/TP rate		Specificity/FP rate		Precision	
			Positive	Negative	Positive	Negative	Positive	Negative
Naïve Bayes (NB)	8479	84.80	0.844	0.851	0.851	0.844	0.856	0.840
Logistic Regression (LR)	*8840*	*88.40*	*0.874*	*0.895*	*0.895*	*0.874*	*0.899*	*0.869*
Decision Trees (DT)	7169	71.70	0.722	0.711	0.711	0.722	0.712	0.722

Stack Ensemble Models		Correct Classification	Accuracy (%)	Recall/TP rate		Specificity/FP rate		Precision	
Base Learners	Meta-Learner			Positive	Negative	Positive	Negative	Positive	Negative
NB/LR	DT	8394	83.94	0.839	0.839	0.839	0.839	0.842	0.837
LR/DT	NB	9040	90.40	0.894	0.915	0.915	0.894	0.919	0.889
NB/DT	*LR*	*9408*	*94.08*	*0.940*	*0.942*	*0.942*	*0.940*	*0.943*	*0.938*

Figure 10: The empirical summary of experimental results of the developed models' classification performances

A. Justification of the validated result based on Accuracy, Efficiency, Class Imbalance and Error Rates

i. Accuracy: Stacked ensembles typically perform better than isolated models. The maximum accuracy of 94.08% was attained by Ensemble 2 (DT and NB as base learners, LR as meta-learner), highlighting the effectiveness of ensemble approaches in enhancing model resilience and generalization.

ii. Error Rate: Stacked ensemble models, especially Ensemble 2, showed a significant reduction in both false positives and false negatives compared to isolated models. This reduction is crucial in applications such as sentiment analysis, since misclassification can have a significant impact on user experience.

iii. Addressing class imbalance: The ensemble approach handled class imbalance better than the isolated model. By combining the strengths of multiple algorithms, the ensemble model was able to balance the decision boundary more effectively and reduce distortions caused by the majority class.

iv. Model efficiency: Although ensemble methods are computationally intensive, the trade-off between improved accuracy and reduced error rate makes them very valuable, especially for large-scale sentiment analysis tasks such as those performed on the IMDB dataset.

B. Histogram Chart Comparing Validated Result of Isolated Models and Stacked Ensemble Models

The below Chart 1(a) and 1(b) Justify the performance classifications of Isolated models and the Stacked-Ensemble models developed for the study.

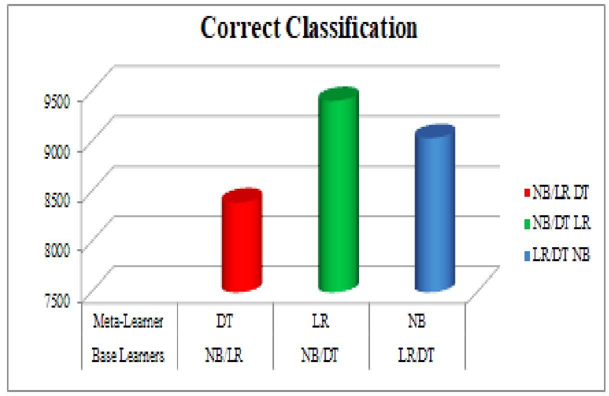


Chart 1(a): Histogram Chart Showing Correct Classification of Isolated Models

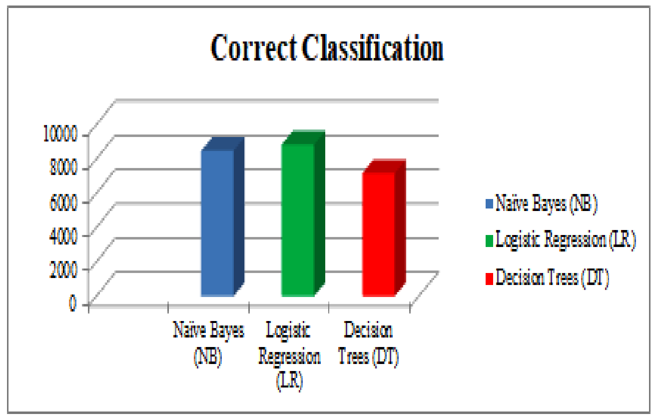


Chart 1(b): Histogram Chart Showing Correct Classification of Stacked Ensemble Models

Comparison of models with highest classification performance and their confusion matrices

Figure 11 (left) shows the confusion matrix results of the LR which has higher performance among isolated learners used in this study. The figure shows that out of 5039 positive reviews, 4532 were classified correctly while 507 were incorrect. Out of the 4961 negative reviews, 4308 were classified correctly while 653 were incorrect. The results indicated that the isolated model of LR had a total of 8840 correct classification owing to accuracy of 88%.

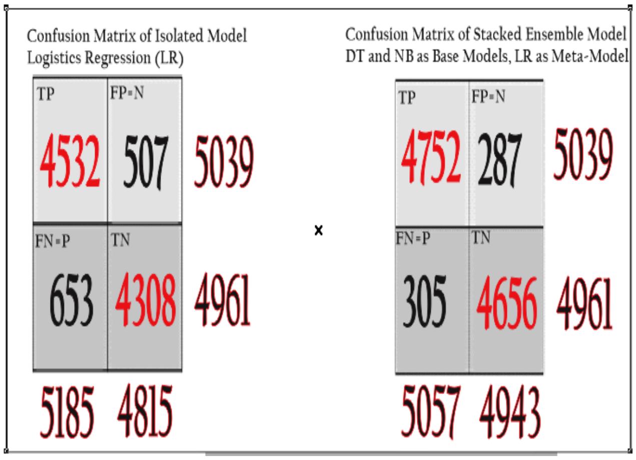


Figure 11: Confusion Matrix of the Stack Ensemble Model with DT/NB as Base Learners and LR as Meta-Learner (left), and Isolated LR (right)

The results support the work of Sulthana, et al., (2022), in opposition to the work of (Luqman, et al. 2023) who stated that NB was a better classifier for sentiments analysis compared to Logistic Regression (LR). Figure 4.9 (right) shows the confusion matrix results of the highest performance stacked ensemble learner which employed LR as the meta-model and DT/NB as base model. The figure shows that out of 5039 positive reviews, 4752 were classified correctly while 287 were incorrect. Out of the 4961 negative reviews, 4656 were classified correctly while 305 were incorrect. The results showed the stack ensemble model that employed LR as the meta-model and DT/NB as base model had a total of 9408 owing to accuracy of 94%. Overall, the aforementioned results showed that using the stack ensemble model instead of isolated classifiers produced better results for sentiment analysis of movie reviews.
The stacking ensemble model with the lowest performance was the one that used DT as the meta-classifier which against the conclusion of (Adeyemi, et. al., 2019) that DT as meta with other stacked algoirthms as base performed excellently better than other stacked models. Apparently, the performance of the one that used LR as a meta-classifier in this study was superior to other utilized models. This support the study of (Gaye et al., 2021), that stacking ensemble in classification performances enhancing high accuracy. Also supported the work of (Basarslan and Kayaalp, 2022), which concluded that ensemble method such as stacking approaches often outperform single algorithm.

Histogram chart comparison of models with high classification performances

Chart 2(a) Shows the Histogram Chart that Compare Correct Classification of Isolated Model and Stacked Ensemble Model and Chart 2(b) Shows the Line Chart that Compare Correct Classification of Isolated Model and Stacked Ensemble Model

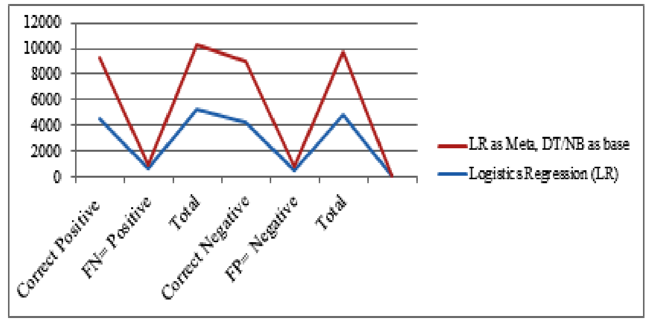


Chart 2(a): Histogram Chart Comparing Correct Classification of Isolated Model and Stacked Ensemble Model of highest result

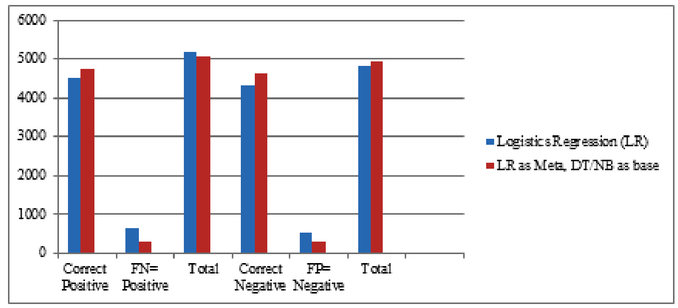


Chart 2(b): Line Chart Comparing Correct Classification of Isolated Model and Stacked Ensemble Model with highest results

Conclusion

This study used a stack-ensemble model comprising three machine learning classifiers to construct a classification model for the automated sentiment analysis of user-provided movie reviews. Among the individual classifiers, LR exhibited the best performance, with an accuracy of 88.4%..The stacking ensemble model significantly outperformed the individual classifiers, achieving an accuracy of 94%. The ensemble's superior performance can be attributed to the combination of the diverse strengths of the base models (DT and NB) with the LR meta-model.

The study came to the conclusion that the number of features that may be retrieved for sentiment analysis from the movie review dataset is further decreased by applying tokenization and stop-word removal with stemming and concludes that adopting Stacking-Ensemble Classifiers can significantly improve sentiment analysis, especially in the movie review context, while addressing the challenges caused by class imbalance and review quality.

The study also underlined how crucial it is to use a stack ensemble of machine learning algorithms rather than standalone ones, which typically have limiting properties. Hence the phrase, two good heads are better than one. The study opens up new avenues for the investigation of ensemble models, particularly the Stacking-ensemble model, which combines many machine learning algorithms with additional sophisticated sentiment analysis methods. This could result in better techniques that are applicable to a variety of fields beyond movie reviews, like public opinion research and product reviews.

References

Adeyemi, O. J., Adebayo, V. O., Olaniyi, O., Olusanya, O., and Idowu, P. A. 2019. "A Stack Ensemble Model for the Risk of Breast Cancer Recurrence," International Journal of Research

Studies in Computer Science and Engineering (IJRSCSE), vol. 6, no. 3, pp. 8–21.

Almi, F. I., and Simatupang, E. C. 2021. "Analysis of Moral Value in 'The Silent Child' Short Movie by Rachel Shenton," Budapest International Research and Critics Institute-Journal (BIRCI-Journal), vol. 5, no. 3, pp. 18 451–18 458.

Baid, P., Gupta, A., and Chaplot, N. 2017. "Sentiment Analysis of Movie Reviews using Machine Learning Techniques," International Journal of Computer Applications, 179:7–17

Basarslan, M. S. and Kayaalp, F. 2022. "Sentiment analysis with ensemble and machine learning methods in multi-domain datasets," Turkish Journal of Engineering, vol. 7, no. 2, pp. 141–148.

Bing Liu (2020). Sentiment Analysis Mining Opinions, Sentiments, and Emotions. 2nd Edition. Cambridge University Press, London.

Bordoloi, M., and Biswas, S. K. 2023. "Sentiment analysis: A survey on design framework,applications and future scopes," Artifcial Intelligence Review, 56:12505–12560.

Brar, G. S., and Sharma, A. 2018. "Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques," International Journal of Applied Engineering Research, vol. 13, no. 16, pp. 2788–12 791.

Danyal, M. M. Khan, S. S., Khan, M., Ghaffar, M. B., Khan, B., and Arshad, M. 2023. "Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques," Journal on Big Data, 5:1–18.

Dashtipour K, Gogate M, Adeel A, Larijani H, Hussain A. 2021. Sentiment Analysis of Persian Movie Reviews Using Deep Learning. Entropy (Basel). 23(5):596. doi: 10.3390/e23050596. PMID: 34066133; PMCID: PMC8151596.

Gan, Z. 2023. "Sentiment Classification of Movie Reviews Based on the Ensemble Machine Learning Model". P. Kar et al. (eds.), Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023), Advances in Computer Science Research 108, https://doi.org/10.2991/978-94-6463-300-9_31

Gaye, B., Zhang, D., and Wulamu, A. 2021. A Tweet Sentiment Classification Approach Using a Hybrid Stacked Ensemble Technique, R. Agerri, Ed.

Hamid A. I. and Abdulazeez A. M. 2024. Sentiment Analysis Based on Machine Learning Techniques: A Comprehensive Review. Indonesian Journal of Computer Science 13(3). DOI:10.33022/ijcs.v13i3.4049.

Horsa, O. G., and Tune, K. K. 2023. "Aspect-Based Sentiment Analysis for Afaan Oromoo Movie Reviews Using Machine Learning Techniques," Applied Computational Intelligence and Computing, pp. 12–12.

Jain, V., Padmavati, M. V., and Roy, P. 2019. "A Brief Survey & Findings on," Trending Issues in Social Media: Sentiment Analysis Approach. IJSRD. International Journal for Scientific Research & Development, 6(1): 1494–1500.

Jim, J. R., Talukder, M. A., Malaka, P., Kabir, M. M., Nur, K., and Mridha, M. F. 2024. "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," Natural Language Processing Journal, vol. 6, pp.

Kaushik, k. and Parmar, M. 2021. "Sentiment Analysis Based on Movie Reviews using Various Classification Techniques," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 7, no. 3, pp. 197–208.

Luqman, M., Yaqoob, A., Ahmad, M. B., and K. Majid, K. 2023. "Sentiment Analysis to Predict Movies Success Rate Based on NLTK Movie Review Corpora Using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(1):46–54.

Mao, Y., Liu, Q., Zhang, Y. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review, Journal of King Saud University - Computer and Information Sciences, Volume 36, Issue 4, 2024,102048, ISSN 1319-1578, https:/ /doi.org/10.1016/j.jksuci.2024.102048. (https:// www.sciencedirect.com/science/article/pii/S131915782400137X)

Mbugua, J., Thiga, M., and Siror, J. 2019. "A Comparative Analysis of Standard and Ensemble Classifiers on Intrusion Detection System," International Journal of Computer Applications Technology and Research, 8(4):107–115.

Sharma, A. and Dey, S. (2013). "A Boosted SVM based Ensemble Classifier for Sentiment Analysis of Online Reviews," ACM Research in Adaptive and Convergent Systems, vol. 13, pp. 43–51.

Sulthana, R. A., Jaithunbi, A. K., Harikrishnan, H., and Varadarajan, V. 2022. "Sentiment Analysis on Movie Reviews Dataset Using Support Vector Machines and Ensemble Learning," International Journal of Information Technology and Web Engineering, no. 1, pp. 17–27.

Wagle, P. P., Manoj Kumar, M. V. 2023. A Comprehensive Review on the Issue of Class Imbalance in Predictive Modelling. In: Shetty, N.R., Patnaik, L.M., Prasad, N.H. (eds) Emerging Research in Computing, Information, Communication and Applications. Lecture Notes in Electrical Engineering, vol 928. Springer, Singapore. https://doi.org/10.1007/978-981-19-5482-5_48