Sentiment Classification of Movie Reviews Based on the Ensemble Machine Learning Model

Zicheng Gan

School of Pharmacy, Fudan University, Shanghai, 200000, China
20211030028@fudan.edu.cn

Abstract. Film reviews play a pivotal role in influencing audience decisions, necessitating accurate classification of sentiments as positive or negative, which holds significant importance for the film industry. To address this, the present study introduces an innovative ensemble learning approach that integrates artificial neural networks, LightGBM, and logistic regression models through a stacking technique. The ensemble model is empirically examined using the IMDB dataset, with a comparative analysis conducted against an individual Artificial Neural Network (ANN) model. The findings demonstrate remarkable enhancements, particularly in terms of accuracy and other relevant metrics, achieved by the ensemble model compared to the individual ANN model, specifically yielding an increase in accuracy from 0.8791 to 0.8904. This substantiates the substantial improvement in accuracy offered by the ensemble model, thereby underscoring the efficacy and potential of ensemble learning for sentiment classification in movie reviews. Moreover, an analysis of the confusion matrix reveals that the ensemble model predominantly improves the classification of reviews labeled as 'positive,' as evidenced by an increase in true positive instances from 4359 to 4451, accompanied by a decrease in false positive instances from 668 to 576. By amalgamating predictions from distinct models, the ensemble model effectively mitigates the limitations inherent to individual models and attains superior performance compared to relying on a single model alone.

Keywords: Machine Learning; Ensemble Model; Artificial Neural Network.

1 Introduction

Film reviews serve as publicly accessible assessments of movies, with dissemination occurring through various channels such as social media platforms and dedicated movie review websites like Internet Movie Database (IMDB) or Douban. The content of these reviews includes, but not limited to, the plot, story content, character portrayals, audio-visual experience, and the performances of the actors. They often incorporate the personal subjective impressions of the viewers during their movie-watching experience, as well as comparisons with other films of the same genre, highlighting noteworthy aspects and areas where the film falls short. Given the internet's evolution and the escalating influence of social media in individuals' lives

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

film reviews have a growing influence on audience decisions regarding which movie to watch in theaters or even whether to go to the movie theater, thereby significantly shaping their initial impressions.

The emergence of movie review websites has revolutionized the way people consume films. Millions of reviews contain valuable insights into audience preferences and feedback which not only play an important part in personal choice. By analyzing these reviews, a plethora of information can be extracted, facilitating more informed business decisions. In terms of sentiment most reviews, they can be divided into three different parts which are positive reviews, negative reviews and ambivalent reviews. Each segment holds significance in its own right, Positive reviews enable the identification of audience-favorite films, facilitating strategic planning for movie promotions and updates to theater schedules. Negative reviews offer crucial feedback on areas of weakness within current movie offerings, guiding efforts to improve future film releases. Additionally, ambivalent reviews shed light on potential sources of confusion or disinterest among audiences, contributing to the refinement of marketing strategies and promotional materials.

Considering the significance of film reviews for individuals and the film industry, along with the varying potential value of different types of reviews, there is a need for an efficient and accurate method for classification. As a classic topic in Natural Language Processing (NLP), there have been several relevant research reports addressing this issue. Some works are related to feature extraction, e.g. Rhetoric Structure Theory tree [1], fuzzy set theory [2] and Gini Index based selection method [3]. Other works are more concentrated on the different algorithms to improve its performance under certain tasks. For instance, Kamal et al. used various supervised algorithms [4], which are Multilayer Perceptron, Naive Bayes (NB), Decision Tree and Bagging for subjectivity and objectivity classification of reviews. Additionally, in Humera Shaziya's research [5], movie reviews are classified by Supporting Vector Machine (SVM) and NB method, and NB model performed much better than SVM in the paper. The study on the data set of twitter posted movie reviews and related tweets have compared different classifiers like SVM [6], Artificial Neural Networks(ANN), NB and k-means clustering algorithm. In another paper [7], Random Forest, NB and K nearest Neighbour models were compared on the reviews collected on IMDB. However, the traditional machine learning methods were not powerful enough. Addressing this task requires advanced NLP techniques and machine learning algorithms.

To address the problem mentioned above, this study proposes to use an ensemble algorithm that combines multiple models for sentiment classification of movie reviews. The study was carried out on the IMDB movie review dataset, and three classifiers: ANN, logistic regression (LR), and lightgbm classifier(LGB) were integrated through a stacking strategy to implement the ensemble algorithm. With each classification algorithm, different feature extraction techniques are applied to the movie review texts, which means the allows for a more comprehensive utilization of data features, thereby overcoming the limitations of a specific algorithm. By comparing the results with the individual ANN, the experiments demonstrate the

effectiveness of ensemble learning with multiple models in improving the performance in terms of sentiment classification.

2 Method

2.1 Dataset Preparation

The IMDB dataset is the Large Movie Review Dataset developed by Standford University[8], which serves as a prominent resource for binary sentiment classification tasks, surpassing prior benchmark datasets in terms of data volume. In total 50,000 reviews, the data is evenly split into each category which accounts for 50%.

First, the text was broken down into individual words through tokenization. Subsequently, Stop Words list was imported to eliminate words that do not contribute much to the sentiment of the review. Finally, before building the Bag-of-Words model for further model training, each sequence of words was padded to the same length as the longest ones. Upon completion of the preprocessing phase, the raw text was represented as a list of indices, with each index corresponding to a specific word. The labeling scheme entailed assigning the value 0 to indicate a negative sentiment and the value 1 to denote a positive sentiment.

2.2 Machine Learning Models

Introduction of ANN. ANN is a machine learning model which mimics the biological structure of neurons and neural networks [9]. Neurons are connected through mathematical expressions to establish nonlinear equations that capture the input-output relationship. This configuration proves to be a powerful tool for solving various problems, particularly in the realms of pattern recognition, image analysis, and speech recognition, and NLP tasks. The ANN mimics the learning and decision-making process of the human brain.

Introduction of LightGBM. LightGBM is a tree-based learning gradient boosting framework [10]. It uses a gradient-based optimization method to build an ensemble of weak prediction models and combines their predictions to make accurate predictions. It is designed to be efficient and scalable, making it particularly suitable for handling large datasets.

Introduction of Logistic Regression. Logistic regression is a frequently utilized statistic model for binary classification tasks. It employs the sigmoid function, compressing numerical values into the range of 0 to 1, to estimate the probability of the input variables.

Ensemble Model. Ensemble learning is a strategy combining multiple individual models to improve the model performance. The ensemble model can capture more diverse patterns and reduce the risk of overfitting, by aggregating the predictions of different models. There are several methods for ensemble learning, and one common approach is stacking, which is a meta-learning ensemble method that involves different base models and then combining their predictions using another model

called a meta-model or stacking model. By combining the predictions of different models through stacking, the ensemble model can take advantage of the strengths of each base model and potentially achieve better performance than using a single model alone.

The process of obtaining new features of base models involves the following steps. Firstly, training multiple base models. The base models are the ANN and the LGB, and each base model is trained on the training set. Secondly, generating predictions. The trained base models are then used to generate a new feature set, where the predictions from the base models are stacked horizontally and each prediction represents a new feature. The stacking model is trained on the new feature set, along with the labels and evaluated on the test set. Once the stacking model is trained, the stacking model is applied to make predictions on the test set which consists of the predictions from the base models. Furthermore, the individual ANN model is trained and evaluated on the same data set.

2.3 Implementation Details

ANN was implemented with Tensorflow V2.12.0 shown in Fig. 1. Input Layer: The model's input is a feature vector of size (9990,). The hidden layers consist of 2 fully connected layers (also known as dense layers) with 128 and 64 neurons respectively. The Rectified Linear Unit (ReLU) activation function is used for these layers. The output layer is a 1 neuron dense layer, employing the sigmoid activation function. The neural network is compiled with the Adam optimizer for parameter optimization. The model's performance is assessed based on the binary cross-entropy loss function and accuracy metric. It is trained for 5 epochs, using a batch size of 64.

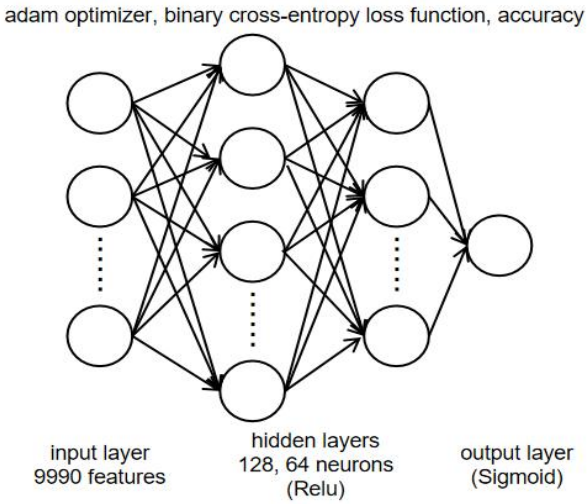

Fig. 1. The structure of ANN model used in this study (Photo/Picture credit: Original).

LGB was implemented with LightGBM (version 3.3.5), where most hyperparameters were set by fault except the number of estimators is set to 500. Additionally, LR was

implemented with scikit-learn (version 1.2.2), where most hyperparameters were set by fault except the max iteration is set to 1000.

3 Results and Discussion

In this study, an experiment was conducted to compare the performance between a single ANN model and an ensemble learning model using the IMDB data set. The ensemble model consisted of ANN and LightGBM Classifier as the base models, with LR serving as the stacking model.

Table 1. The performance of ANN and ensemble model evaluated by various metrics**.**

Model	Accuracy	Precision	F1 score	Recall
ANN model	0.8791	0.8896	0.8782	0.8671
Ensemble model	0.8904	0.8954	0.8904	0.8854

As illustrated in Table 1, the performance was evaluated on four metrics, accuracy, precision, F1score and recall. Based on the results, the single ANN model demonstrated impressive performance in classifying sentiment of movie reviews in the IMDB data set. However, the ensemble learning model proved to be superior in performance compared to the ANN model in all evaluated metrics.

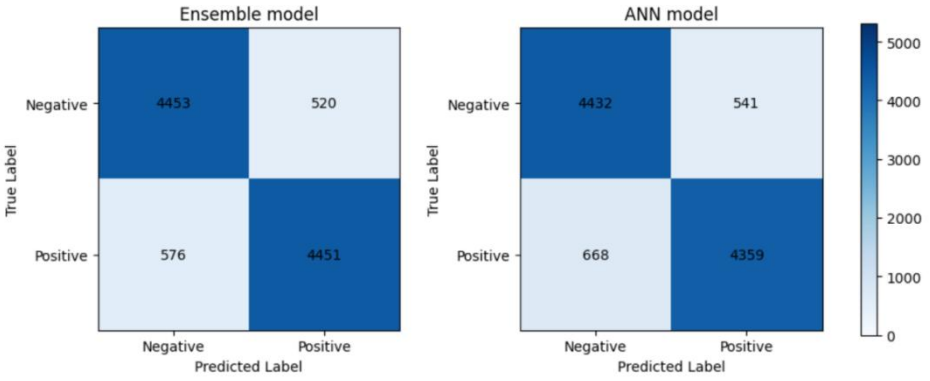

Fig. 2. The confusion matrix of ANN and ensemble model (Photo/Picture credit: Original).

The confusion matrix can facilitate in further analyzing the model's classification outcomes, as shown in Fig. 2. In terms of the ANN model, the confusion matrix reveals how well the model performed in classifying true positives and true negatives. For instance, the model correctly classified 4432 true negatives as negatives but misclassified 541 true negatives as positives. Similarly, the model correctly classified 4359 true positives but misclassified 668 true positives. The confusion matrix for the ensemble learning model demonstrated better performance in classifying true positives and true negatives compared to the ANN model, especially in the

classification of positive data, where there ' s significant difference between the number of true positive and that of false positive.

Overall, these results demonstrate the effectiveness of ensemble learning with multiple models in improving sentiment classification. The ensemble learning model combines predictions from multiple base models, capturing diverse patterns and learning from different perspectives. This can lead to improved accuracy and generalization, as the ensemble can compensate for the weaknesses of individual models. In the results, the main improvement lies in the classification of positive class data.

4 Conclusion

In this study, a novel ensemble learning approach is introduced for predicting emotions in film reviews. The method combines ANN, LGB, and LR models using a stacking strategy. Initially, ANN and LGB are trained separately on the same data set to generates new features, which are then combined using the LR algorithm. The results of the ensemble model demonstrate significant enhancements over the individual ANN model in the conducted experiments.

The ensemble model consistently outperforms the individual ANN in all test metrics, particularly excelling in accuracy. Through analyzing the confusion matrix of both models, great improvement can be seen on the 'positive' labeled data. The reason behind this improvement can be attributed to the ensemble approach's exploitation of the synergies between diverse models. By leveraging the complementary nature of different models, their respective limitations can be mitigated, culminating in an overall performance boost. This underscores the efficacy and promise of ensemble learning in sentiment analysis of film reviews.

The study underscores the potential of combining models with distinct strengths and weaknesses to yield superior outcomes. This approach capitalizes on the unique characteristics of each model to create a more robust and accurate prediction system. In essence, the ensemble strategy proves to be a powerful technique for enhancing the sentiment classification of movie reviews.

However, there are some limitations in this study, only a few algorithms are used in ensemble models. In the future maybe, it can be done on more advanced algorithms, and not only limited to the classification algorithms, but also combined with different feature extraction models, to further investigating the improvement strategy.

References

1. Hogenboom, A., Frasincar, F., De Jong, F., Kaymak, U.: Using rhetorical structure in sentiment analysis. Communications of the ACM, 58(7): 69-77 (2015).
2. Yazdavar, A.H., Ebrahimi, M., Salim, N.: Fuzzy based implicit sentiment analysis on quantitative sentences. arXiv preprint arXiv:1701.00798, (2017).

3. Manek, A.S., Shenoy, P.D., Mohan, M.C.: Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 20, 135–154 (2017).
4. Kamal, A.: Review mining for feature based opinion summarization and visualization. arXiv preprint arXiv:1504.03068, (2015).
5. Shaziya, H., Kavitha, G., Zaheer, R.: Text categorization of movie reviews for sentiment analysis. International Journal of Innovative Research in Science, Engineering and Technology, 4(11): 11255-11262 (2015).
6. Le, B., Nguyen, H.: Twitter sentiment analysis using machine learning techniques, Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015. Springer International Publishing, 279-289, (2015).
7. Baid, P., Gupta, A., Chaplot, N.: Sentiment Analysis of Movie Reviews using Machine Learning Techniques. International Journal of Computer Applications. 179, 45-49 (2017).
8. Maas, A., Daly, R.E., Pham P.T., Huang Dan., Ng A.Y., Potts C.: Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (2011).
9. Agatonovic-Kustrin, S., Beresford, R.: Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5): 717-727 (2000).
10. Ke, G., Meng, Q., Finley, T, et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30 (2017).