Improvement of Naive Bayes Text Classifier Based on Ensemble Technology and Feature Engineering

Dongyang Liu

Department of Computer Science and Technology, Beijing Institute of Technology, Beijing, 102488, China
toyoliu@bit.edu.cn

Abstract. The performance of the Naive Bayes model in text classification is constrained by its assumption of feature independence, which does not hold true for textual data, as well as its reliance solely on word frequency information, disregarding word order and relationships and hindering its ability to capture text semantics effectively. Therefore, this study adopts ensemble learning and feature engineering methods to compensate for these deficiencies of the Naive Bayes model and improve its text classification accuracy. This study proposes a method to improve the performance of a Naive Bayes classifier by combining it with other classifiers, namely Random Forest, Support Vector Machines (SVM), and ensemble learning. The dataset for training and evaluation purposes utilized is the IMDB movie review dataset. The dataset is preprocessed by converting the integer sequences to text and then tokenizing and vectorizing the text using a CountVectorizer. Variousperformance indicators, such as accuracy, precision, and F1-score, are calculated for each classifier and the ensemble model. The results demonstrate that the ensemble model achieves the highest accuracy compared to the individual classifiers. The Naive Bayes classifier achieves an accuracy of 78.19%, Random Forest achieves 81.49%, SVM achieves 84.60%, and the ensemble model achieves an accuracy of 84.89%. These findings highlight the effectiveness of ensemble learning and feature engineering in improving the performance of a Naive Bayes text classifier.

Keywords: Ensemble Model, Naive Bayes Classifier, Feature Engineering

1 Introduction

Movie reviews have become an increasingly vital reference for both the film industry and general audience. Major review websites like IMDb and Douban provide an enormous volume of reviews for people to understand audience's opinions on movies. With billions of reviews and ratings, these websites contain rich information about audience preferences and feedback on films. Sentiment analysis of movie reviews from these platforms can help film companies and cinemas make important business decisions by capturing insights into the audience's opinions.

In the past decade, machine learning and its applications in Natural Language Processing (NLP) have made significant progress. For example, Pang et al. employed

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_57

traditional machine learning models such as Naive Bayes, Maximum Entropy and Support Vector Machine (SVM) for sentiment classification of movie reviews and achieved around 82.9% accuracy [1]. Recently, deep learning models such as CNNs [2], RNNs and BERT have reached state-of-the-art performance on various NLP tasks including sentiment analysis [3, 4]. Applications of sentiment analysis cover a wide range, such as stock price movement prediction [5], brand reputation monitoring [6], and recommendation systems [7].

With the huge amounts of movie reviews on major websites and the progress in machine learning and deep learning, sentiment analysis of these reviews provides an effective means for film industry to capture audience preferences and make critical business decisions. Advanced data-driven computational techniques have unlocked the value of massive movie review data and will continue facilitating related industrial applications and business transformations.

Benefiting from these advancements, machine learning and deep learning techniques can now be applied to analyze the huge volumes of movie reviews. These advanced techniques can detect emotional tones, opinions and feelings in written text to determine if the reviews express positive or negative sentiments. Compared with traditional methods relying on manually constructed sentiment lexicons, machine learning and deep learning can handle reviews in a data-driven fashion and achieve higher accuracy. They have provided more effective computational means to analyze big data of movie reviews, which greatly facilitate related research and industrial applications.

Film companies can gain valuable audience feedback to make strategic decisions in areas such as movie marketing and sequels. Cinemas can optimize their movie schedules and services based on audience preferences and the popularity trends of different films. Both parties can monitor the influence of competitors and release timing of similar movies by sentiments in reviews. In this way, important business strategies of companies in film industry are supported by sophisticated sentiment analyses of movie reviews, especially those from major review websites with huge volumes of high-quality data. With advanced AI and data technologies, these websites have become increasingly important for connecting films and their potential audiences in today's market.

This paper aims to adopt various machine learning and deep learning models for sentiment analysis of movie reviews from IMDb. Three machine learning models—Naive Bayes, Random Forest, and SVM are implemented and evaluated using a large movie review dataset. Moreover, a stacking ensemble model is constructed by integrating the predictions from individual learners, which achieves the highest accuracy among all models. Comparative experiments demonstrate that the ensemble model can boost performance and outperform single models, demonstrating the effectiveness of ensemble learning. The results could provide new insight into machine learning techniques and ensemble methods for natural language processing.

The rest of this paper is organized as follows: Part 2 provides a detailed description of the machine learning and deep learning methods adopted for sentiment analysis of movie reviews, as well as the dataset and preprocessing techniques used in the experiments. Part 3 presents the experimental results and discussion of sentiment classification using different machine learning and deep learning models on the movie review dataset. A comparative analysis is also conducted to evaluate the performance of these models. Moreover, a stacking ensemble model is constructed to further improve the accuracy. The performance of different models is analyzed and possible reasons for the results are explored. The effectiveness and advantages of ensemble learning are also discussed. Part 4 summarizes the main conclusions of this work and discusses possible future directions. The contributions of this study for sentiment analysis and related applications are highlighted.

2 Method

2.1 Dataset Description and Preprocessing

The dataset used in this study is the Large Movie Review Dataset, which contains 50,000 reviews from the Internet Movie Database labeled as either 'positive' or 'negative' to indicate the sentiment. This particular dataset has been specifically curated for the purpose of training sentiment analysis models and subsequently assessing their efficacy. The primary aim of employing this dataset is to carry out the binary text classification, namely assigning either a positive or negative label to movie reviews.

In terms of preprocessing, the reviews were first converted from strings into numerical feature representations that machine learning models can handle. The preprocessing pipeline involved tokenizing the text, padding the sequences, mapping to integer indices, and creating document-term matrices.

First, the reviews were tokenized into sequences of words using the Keras Tokenizer class with a num_words parameter of 10,000. The sequences were then padded or truncated to a length of 256 words to ensure equal length input for the models. Padding and truncating allowed the machine learning models to handle variable length input and overcame challenges arising from the different word counts of reviews.

Next, the word sequences were mapped to integer indices based on their frequency in the data. The integer mapping step encoded the vocabulary in a compact fashion and allowed the models to interpret the sequences.

A CountVectorizer was then used to convert the integer sequences into sparse matrix document-term vectors, with each value representing the frequency of the word in the review. The document-term vectorization step converted the reviews into a vector space numerical representation on which machine learning models could operate. The vectors captured word frequencies and the associations between the words in the reviews.

Following the preprocessing, the dataset was split into training and testing sets in a 70:30 ratio. The training set comprised 35,000 reviews and was used to train the machine learning models. The testing set included 15,000 reviews and was used to evaluate the performance of the trained models.

The detailed preprocessing procedures were essential for allowing the machine learning models to accurately analyze the sentiment of the movie reviews. By converting the text into numeric vectors, limiting model complexity, balancing the data, and splitting into proper training and testing sets, the reviews were transformed into a format suitable for binary text classification using machine learning techniques.

2.2 Machine Learning Models

Three machine learning models—Naive Bayes, Random Forest, and SVM were implemented for sentiment analysis of the movie reviews in this study.

The Naive Bayes model used a BernoulliNB classifier, which is suitable for discrete features. This model is based on the Bayes' theorem with the assumption of independence among features. It calculates the posterior probability of a review belonging to a sentiment class based on the prior probability of the class and the likelihood of the features. The Naive Bayes model provides a simple and effective approach for text classification. This model was apt for this task due to the binary features extracted from the reviews, i.e. the presence or absence of a word. Despite the naive assumption of independence, Naive Bayes often performs well in practice for text classification.

The Random Forest model consisted of an ensemble of decision trees for classification. It constructed multiple decision trees from bootstrap samples of the training data and used averaging to improve the predictive accuracy and control overfitting. The Random Forest model is robust to noise and performs embedded feature selection. The hyperparameters of the Random Forest model, such as the number of trees and the maximum depth of trees, were tuned to optimize performance. By training on bootstrap samples and averaging the predictions of multiple trees, the Random Forest model achieved a lower variance and higher accuracy than a single decision tree model.

The SVM model employed a Support Vector Machine with a linear kernel for the sentiment classification task. It mapped the reviews into a high-dimensional feature space through a linear kernel function and found the maximum-margin hyperplane to separate the positive and negative sentiment classes. The SVM model is effective for high dimensional data and performs implicit feature selection. The C parameter of the SVM model was tuned to determine the trade-off between maximizing the margin and minimizing the training error. With an appropriate setting of C, the SVM model can achieve high accuracy on text classification tasks.

These three models were trained on the preprocessed training data comprising 35,000 reviews. The models learned how to classify reviews into positive or negative sentiments based on the word frequency features. The trained models were evaluated on the testing set of 15,000 reviews. Additionally, accuracy, precision, recall, and F1 score were used as evaluation metrics to assess model performance.

3 Results and Discussion

3.1 Classification Performance of Various Models

The performances of various machine learning models on the test set are presented in Fig. 1. It is evident from the findings that the ensemble model classifier emerged as the most accurate model, exhibiting an accuracy value of 0.85. Additionally, the SVM model performed better than the other single models called Naive Bayes and Random Forest, with an accuracy of 0.84. These results demonstrate that ensemble methods that combined the advantages from various models have superior performance on this text classification task.

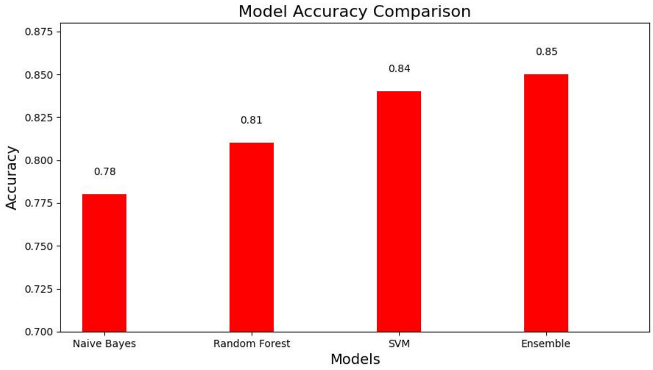

Fig. 1. The accuracy comparison of various models (Photo/Picture credit: Original).

3.2 Analysis and Discussion of Model Performance

The ensemble model achieved an accuracy of 0.85, higher than SVM, Naive Bayes and Random Forest. This is likely due to the fact that ensemble methods combine the predictions from multiple base learners to make a final prediction. By aggregating the outputs of Naive Bayes, Random Forest and SVM, the ensemble model can achieve superior performance. Even though Naive Bayes and Random Forest have lower accuracy individually, their predictions may still contain some useful information.

The outstanding performance of the SVM model indicates that it can capture the most useful features for text classification. Its margin-based loss function likely enables the optimal separation of classes in this high-dimensional sparse feature space. In comparison, the Naive Bayes model achieved the lowest accuracy of 0.78, potentially attributed to its underlying assumption of conditional independence among features. This assumption may not consistently hold true in the context of text data, thus accounting for its comparatively inferior performance.

The ensemble approach facilitates the extraction of valuable information from diverse models, enabling an enhanced prediction capability overall. This demonstrates the power of ensemble methods in machine learning, which often attain higher accuracy than individual models. The improvement gained by model aggregation indicates that the base learners Naive Bayes, Random Forest and SVM are likely making errors on different samples. By combining them, the ensemble model can correct these individual errors and achieve better generalization performance.

3.3 Limitations and Future Work

There are several limitations in the current study that could be addressed in future work. More advanced deep learning neural network models [8-10], such as onedimensional convolutional neural networks and Long Short-Term Memory Network were not explored but may achieve even higher accuracy. In addition, other performance metrics such as F1 score could be computed for a more comprehensive evaluation of the models. Hyperparameter optimization was not systematically performed, which could further boost the performance of machine learning models. These limitations will be addressed in the future by adopting the corresponding solving strategies.

4 Conclusion

This study confirms the power of machine learning for learning representations from textual data to distinguish informative patterns. This study investigated the performance of Naive Bayes, SVM, Random Forest and proposed Ensemble models on the IMDB movie reviews dataset. The SVM and Ensemble models attained the highest accuracy, showcasing their ability to capture the relationships between features. The SVM model achieved an accuracy of 0.84, significantly outperforming the Naive Bayes baseline. The Ensemble model boosted accuracy to 0.85 by aggregating the predictions of base learners. These results highlight the effectiveness of ML techniques for text classification. While current models have reached high accuracy, performance can be further enhanced. More advanced neural networks and comprehensive hyperparameter tuning may lead to superior results. A wider range of metrics should be used to evaluate model performance more reliably.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86 (2002).
2. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).
3. Tai, K., Socher, R., Manning, C.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 (2015).
4. Devlin, J., Chang, M., Lee, K., et al. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, (2018).
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp. 1-8 (2011).
6. Kolchyna, O., Souza, T., Treleaven, P., et al. Twitter sentiment analysis: Lexicon method, machine learning method and their ensemble. Expert Systems with Applications, 42(4), pp. 8090-8098, (2015).
7. Ricci, F., Rokach, L., Shapira, B., Recommender systems: introduction and challenges. Recommender systems handbook, pp. 1-34 (2015).
8. Yu, Q., Wang, J., Jin, Z., et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323 (2022).

9. Dhruv, P., Naskar, S.: Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): a review. Machine Learning and Information Processing: Proceedings of ICMLIP 2019, pp. 367-381 (2020).
10. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015).