ISSN 2959-6157

Sentiment Analysis for Film Reviews Based on Random Forest

Dongling Zheng[1,*]

[1]College of Information Engineering, Sichuan Agricultural University, Yaan, China
*Corresponding author: 202103808@stu.sicau.edu.cn

Abstract:

Sentiment analysis of film reviews has been a popular research topic, and previous researchers have investigated it on the IMDb dataset using a variety of machine learning models, however, the classification results are not satisfactory. Therefore this study aims to construct an effective sentiment analysis model and explore whether the Random Forest algorithm can be applied to the task of sentiment analysis on the IMDb dataset. In this study, after preprocessing the data, the Random Forest model was trained using a training set and evaluated using a test set to explore the accuracy and performance of the Random Forest model in film review sentiment analysis. The study also plotted word clouds to visualize the decision-making effect of the model. The Random Forest Model achieves an impressive 86% accuracy in sentiment analysis, while the word cloud plots provide a visually appealing depiction of its classification task. This indicates that the Random Forest model performs well in the film review sentiment analysis task with high accuracy and performance.

Keywords: Sentiment analysis; Random forest algorithm; IMDb film reviews

1. Introduction

In the modern digital age, the decision-making process of moviegoers is significantly influenced by film reviews, which play a crucial role in shaping their choices[1]. With the boom of the film market in recent years, the huge number of film reviews has made manual analysis and processing impractical, and automated sentiment analysis methods have become an important tool in the film industry and social media, among which sentiment analysis methods using machine learning algorithms have received widespread attention [2]. Sentiment analysis applied to film reviews facilitates the quick and accurate identification of emotional tendencies towards a film, offering valuable insights into the evolving emotions associated with different films.

Currently, Many scholars have done a lot of research on film review sentiment analysis. However, their results are not very satisfactory. The Naive Bayes classifier used by Lopamudra Dey et al. achieves only 82% accuracy [3]. The ComplementNB(Complement Naive Bayes) model used by Christine Dewi & Rung-Ching Chen achieved only 75% accuracy on the task of sentiment analysis on the IMDb dataset [4]. The Support Vector Machine (SVM) model used by Nur Ghaniaviyanto Ramadhan & Teguh Ikhlas Ramadhan achieved only 79% accuracy [5]. The classification results of all the above models are unsatisfactory. Therefore, this study proposes to construct an effective sentiment analysis model using the Random Forest algorithm to obtain better classification results and achieve higher accuracy. Random Forest is an integrated learning algorithm that improves prediction performance by combining multiple decision tree models with high accuracy and robustness [6].

The IMDb dataset used in this study is a widely used online film database [1]. The IMDb dataset is large, covering multiple eras and different genres of movies, and contains a large amount of information about movie reviews, which can be used to research sentiment analysis of movie reviews. In this paper, we use the IMDb dataset to construct a sentiment analysis model and analyze the preferences and trends of users in movies.

This study will also use the visualization method of word cloud mapping to graphically present the high-frequency words in the test set of movie reviews after model training as a way to more intuitively understand the decision-making effect of the model [7].

2. Methods

The Random Forest algorithm-based sentiment analysis model proposed in this paper is shown in Figure 1. As shown in the figure, the model contains five main modules.

Fig. 1 Process of the proposed model

2.1 Data preprocessing

The original dataset contains 25K comments and the corresponding sentiment polarities. Before performing sentiment analysis, data needs to be preprocessed to clean and prepare the data so that it can be better applied to the sentiment analysis model [8]. For the IMDB dataset, the first step is to remove columns in the dataset that have no practical significance for the sentiment score information task. Next, a textual content-based de-duplication operation is performed to avoid duplicate comments from having a repetitive impact on the sentiment analysis results. This is followed by a cleaning and normalization operation that removes special characters, punctuation marks, and numbers from the comment text, retaining only lowercase letters and spaces, which can help to better focus on word-level sentiment analysis [9]. The deactivated words were then removed from the comment text using the list of deactivated words provided by the NLTK library. Discontinued words are common words with no real meaning, such as “a”, “an”, “the”, etc., which are not helpful for sentiment analysis and may interfere with the performance of the model. The comment text is then subjected to a segmentation operation that splits the text into individual words. In addition, using the lexical annotation feature in the NLTK library, lexical labels are added to each word to help better understand the syntactic and semantic roles of words in sentences. Finally, two additional columns were introduced. The first column captures the count of comments in which each participle appears, while the second column records the position index of each participle within the comments. An example of a portion of the pre-processed dataset is shown in Table 1 below.

Table 1. Example of cleaned dataset

Index	Index_content	word	nature	Content_type	Index_word
0	1	love	NN	0	1
1	1	sci	NN	0	2
2	1	willing	JJ	0	3
3	1	put	NN	0	4

2.2 Vectorisation and Data Division

The vectorization method used in this paper is based on the Bag-of-Words Model (BWM). Bag-of-words models are relatively simple to represent and still achieve good results in many sentiment analysis tasks [10]. Each comment is represented as a feature vector, where each dimension corresponds to a word, and the number of times the word appears in the comment is recorded. Each word is mapped to a unique integer index by constructing a vocabulary. Then, for each comment, the number of occurrences of each word in that comment is counted and transformed into a vector representation. This results in a feature vector with dimensions of the size of the vocabulary, where each dimension corresponds to a word. The dataset is usually divided into a training set and a test set. In this paper, the dataset is divided into an 80% training set to train the model and a 20% test set to validate the model performance.

2.3 Random Forest Algorithm

Random Forest is an integrated learning method for solving classification and regression problems. As shown in Figure 2, it makes predictions by constructing multiple decision trees and combining their results to arrive at a final prediction [11]. For the classification problem, the random forest uses voting, where the predictions of each decision tree are statistically voted on and the category with the most votes is selected as the final prediction [12]. The dataset selected in this paper is large, and overfitting

problems may occur if a single decision tree is used. For large-scale datasets and high-dimensional feature spaces, random forests also show better performance. In this paper, we specify the number of parameters to represent the decision tree and then perform a grid search tuning on the training set [13] to evaluate the performance of the model under each combination of parameters by traversing all possible combinations of the given parameters and selecting the best performing combination of parameters as the final model configuration. Finally, the optimal hyperparameters obtained are used to construct the Random Forest model, to obtain the optimal Random Forest classifier.

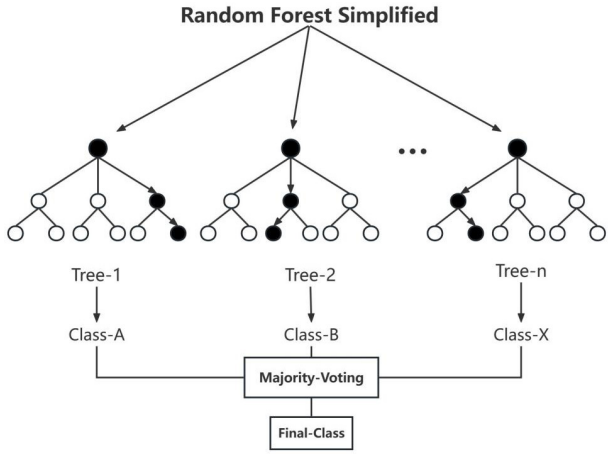

Fig. 2 Random forest Instance

2.4 Model accuracy validation

Accuracy is one of the most commonly used model performance metrics, especially for binary classification problems. The precision of a model is the proportion of correct predictions made by the model across all samples, usually expressed as a percentage. The calculation of accuracy is shown in equation (1) [14].

$$Accuary = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (1)$$

The term TP (True Positive) represents the count of correctly identified positive cases, meaning the cases that are accurately predicted as positive. FN (False Negative) corresponds to the count of cases that are incorrectly identified as negative, indicating the instances falsely classified as negative. FP (False Positive) denotes the count of cases that are mistakenly identified as positive, signifying the instances falsely classified as positive. TN (True Negative) represents the count of correctly identified negative cases, indicating the instances accurately predicted as negative. These four data points can be consolidated and presented in a confusion matrix, as illustrated in Table 2.

Table 2. Confusion matrix

	Negative	Positive
Negative	FN	TN
Positive	FP	TP

2.5 Word Cloud Map

Word cloud mapping is a visualization technique used to show the frequency or importance of different words in textual data [15]. It does this by arranging words in order of their importance or frequency and presenting them in a visually appealing way. Word clouds are usually laid out in a tiled layout where larger words indicate their higher importance or frequency in the text, while smaller words indicate their lower importance or frequency [16].

Some film-related words, such as 'movie' and 'film', may appear more frequently in the text, but may not be very helpful in understanding the specific content of the text. Therefore, in this paper, we choose to eliminate these words when constructing the word cloud map, thus reducing distractions and focusing on the theme and key content of the text [17]. This study visualizes the results of the decision-making of the model through word cloud mapping, which allows for a better explanation of why the model makes a particular classification decision by displaying the keywords in the text that are classified into different sentiment categories. Word cloud maps for different emotion categories are also compared to demonstrate the ability of the model to recognize different emotions.

3. Experiments

3.1 Model performance

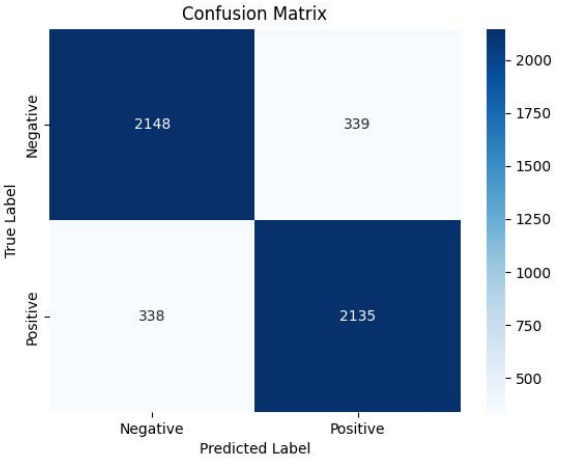

Fig. 3 Confusion matrix result

Figure 3 illustrates the confusion matrix of the model, revealing an accuracy of 86% calculated using Equation (1). The precision, recall, and F1-score metrics, derived from the same confusion matrix, also exhibit a consistent value of 86%. These results collectively indicate that the model

demonstrates a stable and balanced performance on the test set.

Table 3 below clearly demonstrates that the accuracy achieved by the proposed model surpasses that of existing models applied to the IMDb dataset. In reference [5], sentiment analysis of user opinions was conducted using star ratings from multiple comments on the IMDB website, utilizing the SVM method, which yielded an accuracy of 79%. Similarly, reference [4] employed the Word Frequency-Inverse Document Frequency Model (TFIDF) for feature selection and extraction, coupled with the ComplementNB classifier, achieving a maximum accuracy of 75%. Furthermore, reference [3] achieved an accuracy of 82% for the classification of film reviews using the Naive Bayes classifier.

In contrast, the Random Forest classifier developed in this study exhibits a significant improvement in accuracy compared to the aforementioned models. These findings substantiate the stable and well-balanced performance of the proposed model in sentiment classification tasks, providing robust support for sentiment analysis of textual data, such as movie reviews.

Table 3. Confusion matrix

Different Model on IMDb Dataset	Accuracy
Random Forest	86%
ComplementNB	75%
SVM	79%
Naive Bayes	82%

3.2 Word cloud mapping

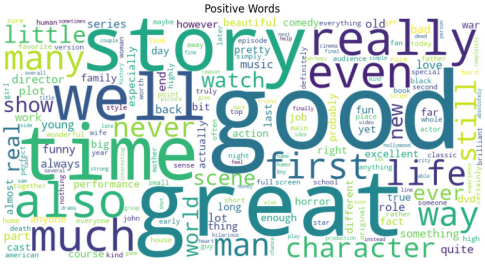

Fig. 4 Positive Comment Word Cloud

Figure 4 shows the word cloud of the positive comment texts decided by the model on the test set, and it can be seen that “great” and “good”, which are two obvious positive sentiment words, appear with the greatest frequency. Also included are “story”, “well”, “time”, and so on.

Fig. 5 Negative Comment Word Cloud

Figure 5 shows the word cloud of negative comments on decision-making, with “bad” and “even” appearing most frequently, along with “time”, “first”, “much”, and so on. The word cloud also shows that the model is more effective in determining the sentiment polarity of the review texts of film reviews. The word clouds of positive and negative reviews contain two words “good” and “bad” that express emotions very directly, respectively.

4. Conclusion

This study aims to explore the application of a sentiment analysis model based on the Random Forest algorithm in the sentiment analysis of film reviews, through an empirical study using the IMDb dataset.

The Random Forest algorithm is an effective classifier for dealing with large-scale datasets and high-dimensional feature spaces. The sentiment analysis model of this Random Forest Algorithm achieved 86% accuracy on the test set by validating the accuracy of the model. This accuracy rate is higher than other common classifiers such as Complement Naive Bayes Classifier and Support Vector Machine (SVM) models, indicating that the model has high performance on the film review sentiment classification task.

In addition, by plotting the word cloud graph, it is observed that the model is more effective in judging the sentiment polarity of positive and negative comment texts. The word cloud map shows that in positive comments, positive sentiment words such as “great” and “good” appear more frequently, while in negative comments, negative sentiment words such as “bad” and “even” are more frequent in negative reviews. This further validates the ability of the model to accurately classify the sentiment of film reviews.

In summary, the sentiment analysis model based on the random forest algorithm shows good performance in film review sentiment analysis. This study provides a useful reference for the further development and application of sentiment analysis models based on the random forest algorithm. However, this study still has some limitations. We only used the random forest algorithm as a sentiment

analysis model, and other machine learning algorithms or deep learning models could be considered for the application. Future research could further explore the combination of different algorithms and datasets to improve the performance and generalization of sentiment analysis models.

References

[1]Mahyarani M, Adiwijaya A, Al Faraby S, et al. Implementation of Sentiment Analysis Movie Review Based on Imdb with Naive Bayes using Information Gain on Feature Selection. IEEE, 2021: 99-103.

[2]Steinke I, Wier J, Simon L, et al. Sentiment Analysis of Online Movie Reviews using Machine Learning. Int. J. Adv. Comput. Sci. Appl, 2022, 13(9): 618-624.

[3]Dey L, Chakraborty S, Biswas A, et al. Sentiment Analysis of Review Datasets using Naive Bayes and K-NN Classifier. arXiv preprint arXiv:1610.09982, 2016.

[4]Dewi C, Chen R C. Complement Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database. Cham: Springer International Publishing, 2022: 81-93.

[5]Ramadhan N G, Ramadhan T I. Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM. Sinkron: jurnal dan penelitian teknik informatika, 2022, 7(1): 39-45.

[6]Genuer R, Poggi J M, Genuer R, et al. Random Forests. Springer International Publishing, 2020.

[7]Wu Shiqi, Zhao Xing-yu, Qiu Fenglin, et al. Analysis of JD Commodity Evaluation Word Cloud Based on Web Crawler. International Journal of Advanced Networking and Applications, 2021, 12(5): 4668-4676.

[8]Aufar M, Andreswari R, Pramesti D. Sentiment Analysis on Youtube Social Media using Decision Tree and Random Forest Algorithm: A Case Study. IEEE, 2020: 1-7.

[9]Maharana K, Mondal S, Nemade B. A review: Data Preprocessing and Aata Augmentation Techniques. Global Transitions Proceedings, 2022, 3(1): 91-99.

[10]Yan Dongyang, Li Keping, Gu Shuang, et al. Network-based Bag-of-words Model for Text Classification. IEEE Access, 2020, 8: 82641-82652.

[11]Han S, Kim H, Lee Y S. Double Random Forest. Machine Learning, 2020, 109: 1569-1586.

[12]Khomsah S. Sentiment Analysis on Youtube Comments using Word2vec and Random Forest. Telematika: Jurnal Informatika dan Teknologi Informasi, 2021, 18(1): 61-72.

[13]Belete D M, Huchaiah M D. Grid Search in Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results. International Journal of Computers and Applications, 2022, 44(9): 875-886.

[14]Qaisar S M. Sentiment Analysis of IMDb Movie Reviews using Long Short-term Memory. IEEE, 2020: 1-4.

[15]Khrais L T. Role of Artificial Intelligence in Shaping Consumer Demand in E-commerce. Future Internet, 2020, 12(12): 226.

[16]Dubey A D. Twitter Sentiment Analysis During COVID-19 Outbreak. Available at SSRN 3572023, 2020.

[17]Wu Yingcai, Provan T, Wei Furu, et al. Semantic-preserving Word Clouds by Seam Carving. Oxford, UK: Blackwell Publishing Ltd, 2011, 30(3): 741-750.