Sentiment Analysis: Predicting the Movie Ratings On Douban by Machine Learning Methods

情感分析：用机器学习的方法预测豆瓣电影评分

Keni Xue
Student ID: 1716395
Supervisor: Tai-Jun Chen

May 4th 2022

Abstract

Currently, watching movies becomes a very popular leisure activities for people. With the development of Internet, more and more people tend to watch the movies on Internet and write down the movie reviews on social media, such as IMDb and DouBan. Meanwhile, potential viewers tend to read the ratings and reviews about movies on these websites before select movies to watch. Moreover, the movie industry becomes an important part of global commerce. Thus, it is important to find out the opinions and sentiment behind the movie reviews. The sentiment of the movie reviews can be identified by sentiment analysis. In this study, sentiment of Chinese movie reviews on DouBan dataset are analyzed and classified into classes on scale of 1 to 5, where the 5 classes from small to large represent extremely negative, negative, neutral, positive and extremely positive ones, respectively. Before training the models, data preprocessing and feature selection are inevitable. Then classification models are created by reviews data using Multinomial Naive Bayes, Random Forest, Support Vector Machine, Logistic Regression and K-Nearest Neighbor. Finally, the Multinomial Naive Bayes method produces the best accuracy(60%), and the logistic regression model is the most robust model with highest F-1 score(40.16%).

Keywords: Sentiment Analysis, Multinomial Naive Bayes, Random Forest,Support Vector Machine, Logistic Regression, K-Nearest Neighbor

Abstract

现在，对人们来说，看电影变成一项越来越受欢迎的休闲活动。随着互联网的发展，人们喜欢在网上看电影，并且喜欢在豆瓣或者 IMDb 这种社交媒体上发表自己的影评。同时，潜在的观众倾向于选择要观看的电影前，浏览该电影的评分和评论。而且电影产业也变成了世界贸易非常重要的一部分。所以发掘电影评论背后的观点和情感变得非常重要。情感分析方法可以用来识别电影评论的情感。在本文中，笔者将分析豆瓣数据集中的电影评论的情感，并且将影评的情感分为 1-5 类。这五类从小到大分别是极其消极，消极，中立，积极和极其积极。在训练机器学习的模型前，需要进行数据预处理和特征挑选。然后，用电影评论的数据集来训练分类模型。本文中，多项式朴素贝叶斯，随机森林，支持向量机，逻辑回归和 K 最临近模型来解决这个多分类的电影评论情感分析问题。最后，多项式朴素贝叶斯模型的精度 (60%) 最好。逻辑回归模型的最鲁棒，因为有最高的 F-1 score(40.16%)。

关键词：情感分析，多项式朴素贝叶斯，随机森林，支持向量机，逻辑回归和 K 最临近

Contents

1 Introduction

Currently, watching movies becomes a popular leisure activity, especially among the young generation. Thus, movies become not only a source of recreation but also a major part of marketing and global commerce. As an important part of global commerce and marketing, the success of movies is deeply concerned by box office officials, movie directors, and general people [1]. Moreover, the ratings and reviews of movies on popular movie review aggregation sites like IMDb (Internet Movie Database), Douban, and Rotten Tomatoes are influential factors in the box office and the success of the movies. It is because the majority of people are accustomed to writing down some reviews and providing feedback on those platforms after watching movies, and potential viewers tend to choose movies to watch depending on reviews and ratings. In particular, potential viewers tend to read the movies' comments and ratings before watching some pay-per-view movies [2]. Generally, viewers are inclined to watch movies with higher ratings and positive feedback. It is certain that some movies' ratings and comments mislead the potential viewers to make the right choices, since the emotions in the ratings and comments sometimes do not match the emotion evoked by the movie itself [3]. Rather, some emotions behind the movie reviews and ratings depend on individual experience and personalities. However, the film reviews and ratings are still crucial for success of movies, because those are the important criteria for people to choose the movies to watch and for directors and actors to shoot the next films.

Douban website is one of largest and most popular film sharing and comment communities in China, where Chinese are accustomed to writing down and reading movies' comments and ratings. This website was founded in 2005 and similar to IMDb. On DouBan, there are millions of information about worldwide films and filmmakers, including movie comments, ratings, film cast, etc. People can search for and determine the movies that they tend to watch by browsing through the site. The reason why DouBan become more and more popular in China is that users think that they can get honest reviews and ratings there [1]. Thus, there are a huge amount of data about ratings of movies and comments on Douban. With the increasing number of movies' reviews and ratings, a single person can't analyze all of the sentiment of users' comments and texts [4]. It is because handling and analyzing the huge amount of text comments is time-consuming and highly cost-oriented [5]. Therefore, sentiment analysis is introduced to analyze the sentiment of the huge amount of movie reviews on social media like DouBan.

On the other hand, sentiment analysis is a classification process to analyze documents' opinions, interactions, and emotions [2]. According to Lima et al., people can understand the relationship between humans emotions and natural text by sentiment analysis [6]. Currently, sentiment analysis has become one of the most popular and trending research fields in text mining and natural language processing (NLP). Thus, the present generation tends to apply the method of sentiment analysis to analyze the user' s reviews about the products, movies, health care systems, etc [7]. The sentiment analysis does benefit many fields. For example, sentiment analysis contributes to success of business, because products' reviews on the Internet are one of the crucial factors leading to the success of the product. [6]. Through analyzing the sentiment behind reviews, marketers can not only find out the basic needs and requirements of the customers but also check the satisfaction of the product quality and pricing plans from the customers, which leads to produce goods with better quality [4]. For movie reviews, sentiment analysis helps producers acquire the characteristics of success of the movies from the movie reviews on movie review aggregation sites. Thus, analyzing the sentiment of a huge amount of the movie reviews on those websites is important. These sentiment analysis can be achieved by machine learning and the latest deep learning algorithms efficiently, which includes the understanding of text processing and analysis [5]. It is because machine learning can process the data at a higher rate in a short period, and then produce the efficient and accurate results [5].

In order to classify the sentiment or opinions expressed in the huge amount of reviews, several types of machine learning methods are introduced. Machine learning is a method for building intelligent algorithms that solve problems by learning from existing data. This is carried out by studying the available data patterns to construct a mathematical model and make a decision [8]. The various machine learning methods can be used to analyze the sentiment of movie reviews. Therefore, in this study, machine learning is implemented to help producers and directors to analyze the sentiment of Chinese movie reviews on Douban. Those machine learning methods try to classify the rating of reviews into classes on a scale of 1 to 5, depending on the sentiment of reviews. The ratings of the movies represent the attitudes about the movies, where ratings from 1 to 5 represent extremely negative, negative, neutral, positive, and extremely positive ones, respectively. This study implements the Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Multinomial Naive Bayes (MNB) to classify the sentiment on movie reviews. Moreover, in order to elevate the performance of each of these baseline classifiers, some form of pre-processing could be employed such as removing stopwords and punctuations. Those models can be evaluated and compared using evaluation parameters, such as accuracy, precision, recall, and F-1 score on the Douban movie dataset.

The rest of the paper is organized as follows: Section II is literature review about sentiment analysis. In Section III, the methodologies related to sentiment analysis are discussed. Section VI presents description of dataset and results of the proposed machine learning models. In section V, drawing some conclusions based on experimental analysis and discussing the future work.

2 Literature Review

Many researchers have used various machine learning methods to research sentiment analysis. In the 1960s, the research on Natural Language Processing (NLP) started when NLP systems SHRDLU and ELIZA was developing. SHRDLU worked with restricted vocabularies and ELIZA simulated a psychotherapist which provided human-like interaction [9]. Furthermore, in 1980, Russell et al. thought emotions had two dimensions, which were activation and pleasure, and then proposed the circumplex model [3]. After that, Stone applied a database that contained a lexicon of emotions with more than 100 categories and 11,000 words to Russell' s model [3]. Sentiment analysis can analyze not only English texts but also texts with other languages, such as Chinese texts. For instance, Prasad et al. analyzed a massive number of Indian comments on Twitter, and then classified those comments into positive, negative, and neutral ones [10]. Similarly, Zhang and Zheng compared performance of different machine learning algorithms when analyzing sentiment of Chinese texts [9].

In the processes of sentiment analysis, feature selection is a important step to improve the performance of models. In 2009, Yessenov and Misailovic proposed different methods to extract text features, such as the bag-of-words model, using a large film review corpus, using WordNet's synonym knowledge, limiting adjectives and adverbs, limiting word frequency with thresholds, and handling negation [2]. Moreover, Ulfa and Irmawati tended to combine Mutual Information feature selection method and the Naive Bayes classification model to build a sentiment analysis system. Then the classification time was decreased by 51.52% and the accuracy was increased by 1.7%, because of the Mutual Information feature selection method [2].

On one hand, there has been a lot of work related to sentiment analysis in the businesses area. For example, Zhao et al. proposed Weakly-supervised Deep Embedding in sentiment analysis of product reviews to determine whether or not to buy the product [7]. Furthermore, Srivastava et al. predicted the sentiment of restaurant reviews based on a subset of the Yelp Open Dataset. In this study, Srivastava et al. used more than ten machine learning and deep learning models to analyze the sentiment of reviews, such as Gradient Boosting, Multilayer Perceptron, Hierarchical Attention Network, etc. Moreover, Srivastava et al. proposed a new Multi-tasked joint BERT model that improved the overall classification performance [11].

On the other hand, lots of studies related to sentiment analysis of movie reviews have been done. In the process of feature selection, Asriyanti et al. used Information Gain method to eliminate features which were not useful for improving accuracy, because Asriyanti et al. thought data problems in movie review led to the slow and less sensitive sentiment analysis [12]. When analyzing the sentiment about movie reviews, many machine learning models are used in reasearches. For example, Ramadhan et al. applied two machine learning classifiers, which were Support Vector Machine (SVM) and Logistic Regression [13]. In this study, the performance of SVM classification was superior to the performance of Logistic Regression. Moreover, Yasen et al. used eight types of machine learning classifiers to analyze the IMDb movie reviews dataset. In this study, the classifiers were Naive Bayes (NB), Random Forest, Decision Tree (DT), Ripple Rule Learning, K-Nearest Neighbours (KNN), Bayes Net, SVM and Stochastic Gradient Descent. According to Yasen et al., the Ripple Rule Learning algorithm gave the worst results, but Random Forest outperformed other classifiers [14]. Similarly, Quader et al. found the relationship between the success rate of movies and sentiment of movie review data using Support Vector Machine and Neural Networks. In this experiment, Quader et al. analyzed the sentiment of reviews on different dataset, such as IMDb, Rotten Tomatoes, and Box office Mojo dataset. Then Quader et al. figured out that the accuracy of those methods were 80-90% [7]. Futhermore, Tripathy et al. did the sentiment analysis with four classifiers, which were Maximum Entropy, SVM, NB and Stochastic Gradient Descent for sentiment analysis on the IMDb dataset. Meanwhile, Tripathy et al. compared the performance of the classifiers when features had different "n-gram" , where an "n-gram" was a connected string of N items from a sample of text in the process of tokenization. Then Tripathy et al. found that the longer "n-gram" it is, the higher accuracy it is [15]. Additionally, Narayanan et al. aimed to improve the performance of the NB classifier by combining negation handling, feature selection, and word n-gram techniques [16]. However, best result of 88.8% has been outperformed in the proposed approach. Besides, through computing sentiment on IMDb movie reviews, Bansal et al. predicted the genre of movies based on users' reviews on Twitter and recommended movies for users depending on their requirements. At same time, Bansal et al. classified the text on a scale of 0 to 4, which was a multi-classfication task based on machine learning [17].

The sentiment analysis are also achieved by some deep learning techniques. For example, Mathapati et al. classified reviews using Convolution Neural Network(CNN), Recurrent Neural Network(RNN), Long Short Term Memory(LSTM) and the model which collaborated Convolution Neural Network and Long Short Term Memory(CNN-LSTM). Then Mathapati et al. found sentiment analysis using CNN-LSTM model had highest accuracy and least loss with a least computing time among those four models [7]. Similarly, Lima et al. analyzed sentiment classification on the IMDb movie reviews by CNN and LSTM models. However, the results had shown that CNN had outperformed LSTM and CNN-LSTM for sentiment classification on IMDb movie reviews [6]. Moreover, in research done by Gogineni et al., Recurrent Neural Network, Naive Bayes, SVM, Random Forest, KNN, Gated Recurrent Units and LSTM were used for sentiment analysis. The result was that Naive Bayes model had best performance in machine learning methods and LSTM model had best performance among deep learning methods in the context of the text. As for the word embedding process, Gogineni et al. compared Term Frequency-Inverse Document Frequency (TF-IDF) method and Continuous Bag of words (CBOW) method with same deep learning models. Gogineni et al. found that sentiment analysis with TF-IDF method had better performance than sentiment analysis with CBOW method when using deep learning algorithms [5].

3 Methodology

3.1 Data preprocessing

Data preprocessing is a series of processes to remove the useless information from the gathered raw text, where the useless information is the information that is harmful to the training process or easily lead to confusion in the classification process [18]. Generally, data preprocessing prepares the data according to the requirements of classification. It is a necessary and important process, because classifiers require data in a prescribed form rather than the form of paragraphs. The data which are not cleaned and organized well are likely to lead false identifications [4]. Hence, data preprocessing is an important task in the process of data mining.

Before analyzing the sentiment of movie reviews on the DouBan movie review dataset, there are several steps for data preprocessing. Firstly, remove all the punctuation such as "!", "? ". Secondly, all the HTML tags are removed from the text. The movie reviews have a lot of HTML tags, because these data are majorly gathered from Internet. Those hybrid links should be scrapped, since those links do not give any insights but become vectors that pose a threat in terms of memory [4]. It is one of the efficient processes to decrease the errors and complexity when predicting the sentiment of reviews. Thirdly, all traditional Chinese characters in the dataset are converted into simplified ones in Chinese. It is because different forms of Chinese characters are treated as different features when training models. The more features in models, the lower the efficiency and accuracy in the experiment of sentimental analysis. For example, "难看" in simplified Chinese characters meaning "awful" is written as "難看" in traditional form. Both words have the same meanings but different forms. Thus, the traditional Chinese characters should be converted into corresponding simplified Chinese characters. Fourthly, all the English words should be removed, and only Chinese characters in reviews are kept and analyzed. It is because a small number of English words increase the features, which increase the complexity of the models and decrease the accuracy when training the models. The fifth step is tokenization, which is a technique to divide a sentence into a stream of tokens known as words, symbols, and meaningful phrases [5]. Tokenization is very important for sentiment analysis, because tokens impart meaning to the statements, and then the implications of the reviews are easily understood [4]. However, the wrong token imparts causes wrong conclusions, so it must be done carefully in the process of tokenization [4]. In this paper, tokens are created by JIEBA, which is a Python Chinese word segmentation library. Sixthly, remove the stopwords in the movie reviews. Stopwords are the words that complete a sentence but have no or negligible emotion of the reviews or texts, such as "me" , "you" , "is" and "the" etc. This step is crucial, because removing these words in the documents increases efficiency when training machine learning models and improves performance in the process of sentimental analysis. Finally, remove missing values in columns labeled "CONTENT" and "RATING" . The missing value is one type of the outliers, which leads to errors when training the machine learning models. The following examples show the processes of tokenization and removing stopwords.

Sentence = I love eating ice cream.

Tokenization = [I, love, eating, ice, cream]

Removing stopwords=[love, eating, ice, cream]

3.2 Vectorization

Vectorization is a process that extracts features from movie reviews and then converts them into an n-dimensional vector. It is one of the main prerequisites for most of the NLP tasks to deal with the natural text [6]. The aim of vectorization is to transform the reviews into a mathematical structure, so that the classifier models can understand and compute those reviews [4]. In this paper, Term Frequency-Inverse Document Frequency (TF-IDF) is picked up as the method of vectorization.

TF-IDF is a statistical method that evaluates the importance of a word to a document or to a corpus. TF-IDF is an approach in Natural Language Processing (NLP) to extract important keywords in reviews or documents.

$$TF - IDF = TF * IDF$$

On one hand, TF is the frequency of a word, which means the number of times that the word appears in the document. The more frequently the term appears, the higher the TF value of the term it is. The TF value normalizes the term count, which prevents the weighting of the given word from being biased by longer documents. On the other hand, IDF is inverse document frequency, which means how common a word is in the document. If the IDF value of a word is closer to 0, the word is more common in the document set. IDF can increase the weighting of infrequent terms, while decreasing the impact of frequent terms. The TF-IDF formula is as follows

$$TF - IDF(D, d) = TF * log(\frac{D}{d})$$

and

$$TF = \frac{n}{N}$$

Where:

D is the number of documents.

d is the number of documents containing the given word.

TF is frequency of the word appearing in a document.

n is the number of given word appearing in a document.

N is total number of the words in a document.

Hence, the higher the TF-IDF value is, the more important the word is. Moreover, when the a term is not important, then TF-IDF value is equal to 0.

3.3 Logistic Regression

Logistic regression is a classification algorithm in machine learning. It is used to solve binary classification problems, such as -1 or 1, and True or False. It is a supervised classification algorithm. Suppose there are two label class $y_i \in \{-1, +1\}$ in the dataset. The hypothesis function of the Logistic Regression is

$$h_w(x) = \frac{1}{1 + e^{-w^T x}}$$

Where:

$h_w(x)$ represents probability that y equals to 1, and is a function of x. Thus $0 \leq \ h_w(x) \ \leq 1$.
$x = [x_1, x_2, \cdots, x_n]$ is the vector of explanatory variables .
$w = [w_1, w_2, \cdots, w_n]^T$ is a vector of weight.
Hence,

$$w^T x = \sum_{i=1}^{n} w_i x_i = w_0 + w_1 x_1 + \cdots + w_n x_n$$

Since Logistic Regression model is a binary problem, the model response is symmetrical shown as

$$P(y = 1|x; w) = h_w(x)$$

$$P(y = 0|x; w) = 1 - h_w(x)$$

Where

$$P(y = 1|x; w) + P(y = 0|x; w) = 1$$

Using maximum likelihood estimation method, cost function of Logistic Regression $J(w)$ is obtained by

$$J(w) = \frac{1}{m} \sum_{i=1}^{n} Cost(h_w(x_i), y_i) = -\frac{1}{m} \sum_{i=1}^{n} [y_i log(h_w(x)) + (1 - y_i) log(1 - h_w(x))]$$

In order to minimize the cost function $J(w)$, “liblinear” method is used to find out hypothesis parameters w. However, Logistic Regression model has some limitations. For example, Logistic Regression is assumed to be linear between the dependent and independent variables. However, in the real world, most of data is a jumbled mess, rather than linearly separable. Hence, most of data in real world is not suitable for Logistic Regression. In this experiment, Logistic Regression method is used by calling LogisticRegression() method on “sklearn” .

In this paper, there are five classes rather than two classes. Thus, the ”One-versus-Rest”(OVR) method is introduced to solve the multi-class classification problems in Logistic Regression. One-vs-rest (OvR) is a method which uses binary classification method to solve multi-class classification problems. It splits a multi-class problem into many binary classification problems. Then a binary classifier is trained for each binary classification problem, and the most confident model is used to make predictions after training all the binary classifiers.

For example, there is a multi-class classification problem, which includes classes 'pink', ‘blue' and 'green'. This multi-class classification problem could be split into three binary classification problems as follows:

The First Binary Classification Problem: pink vs [blue, green]

The Second Binary Classification Problem: blue vs [pink, green]

The Third Binary Classification Problem: green vs [pink, blue]

Then train those three standard Logistic Regression classifiers $h_w^i(x)$ for $i \in \{-1, +1\}$ to find out the decision boundary. However, there are several drawbacks of "OvR" method. For example, each class requires to create one model. In this experiment, five classes require five models. However, if there is a large datasets with huge amount of data or hundreds of classes, OvR method is not suitable.

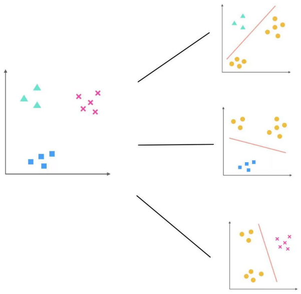

Figure 1: One-Versus-Rest Method

3.4 Support Vector Machine

Support Vector Machine (SVM) is a commonly used machine learning method to solve binary classification problems. The idea of SVM classification method can be described as follows. Suppose there are a training set $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, where $y_i \in \{-1, +1\}$ is the coded label class and $x = (x_1, x_2, \cdots, x_n)^T$ is the n-dimensional feature of the sample. y_i is equal to +1, when sample x_i is assigned to positive class; and y_i is equal to -1, when sample x_i is assigned to negative class. Define a hyperplane as follow:

$$w^T x + b = 0$$

This hyperplane can divide the training set. $w \in R^n$ is a vector of the weight which is orthogonal to the hyperplane and $b \in R^n$ is the bias distanced from the origin to the hyperplane. The aim of SVM is to maximize limit between classes or margin of an separating hyperplane [8]. Hence, the problem of SVM optimization is

$$min\frac{1}{2}\|w\|$$

Using linear kernels to train this model and find suitable hypothesis parameters. However, SVM requires very high memory and the algorithmic complexity. It is bacause, on SVM model, all support vectors are stored in memory and this number increases suddenly with the training dataset size. Thus, the computation time of SVM algorithm is very slow when training the model. The SVM method is used by calling SVC() method of ”sklearn”.

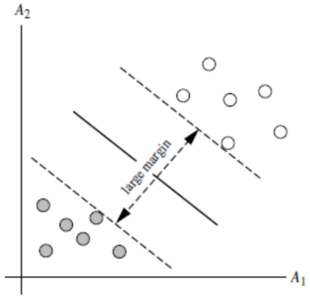

Figure 2: Support Vector Machine

In this paper, there are five class, which means $y_i \in \{1, 2, 3, 4, 5\}$. For this multi-class classification, the OVR technique mentioned in Logistic Regression is used to solve the multi-class classification problem in SVM.

3.5 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a machine learning model which is used to classification problems. KNN algorithm was introduced by Fix and Hodges and developed by Cover and Hart [8]. The KNN predicts the label on the test sample with the majority label from its training sample with the nearest distance or nearest neighbor [8]. The KNN method is that, according to the the distance between samples, the data on dataset are classified. In this method, the Euclidean distance is introduced, because Euclidean distance is convenient, efficient and productive to apply. The formula for calculating the proximity as follows:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The range of proximity value is between 0 and 1. When proximity is equal to 0, then the two absolute cases are different. Inversely, when proximity is equal to 1, two absolute cases are extremely similar. However, KNN is sensitive to noisy data, so it is necessary to manually impute missing values and remove outliers. The KNN algorithm is used by calling KNeighborsClassifier() method of ”sklearn”.

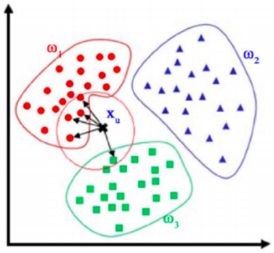

Figure 3: K-Nearest Neighbor Method

3.6 Random Forest

Random forest is a very common method in machine learning for the classification problem, which is an extension of the Decision Tree. The Random Forest is a set of classifiers consisting of multiple decision trees, and the classification result depend on the classification results of each member decision tree. There are several steps on Random Forest algorithm. Firstly, n number of training samples are randomly chosen from the original training set, and then build n random decision trees. Secondly, after constructing the decision trees,voting are done in each class from the sample data. Thirdly, votes from each class are combined and then the most votes are taken. Then the best vote are produced by using random forest in the data classification. However, Random Forest still have some limitations. For example, Random Forest algorithm requires much more resources and computational power, because it builds a lot of decision trees in the process of training data. The Random Forest algorithm is used by calling RandomForestClassifier() method of "sklearn".

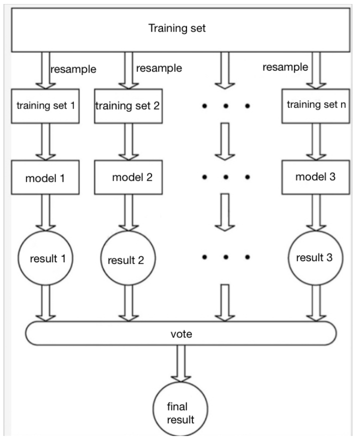

Figure 4: Random Forest

3.7 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) algorithm is a commonly used machine learning method to solve multi-class classification problems. This algorithm is based on Bayes' Theorem. MNB is derived from Naive Bayes. MNB classifier has many advantages. For example, it is very fast, simple and easy for MNB classifier to implement. Moreover, it requires less training data to train the classifier and is highly scalable [7]. Suppose there are a training set $(x^1, y_1), (x^2, y_2), \cdots, (x^n, y_n)$, where $y_i \in Y = \{1, 2, 3, 4, 5\}$ is the coded label class and $x^i = (x_1^i, x_2^i, \cdots, x_m^i)^T$ is the m-dimensional feature of the sample i. x^i represents a movie review in dataset, and x_j^i represents word j in review x^i. Thus, calculating the probability of a document x^i belonging to class $c \in Y$, formula is below:

$$P(x^i|c) \propto P(c) \prod_{j=1}^{m} P(x_j^i|c)$$

Where:

$P(x^i|c)$ is the probability of a review x^i in the class c.

$P(c)$ is the prior probability of a class c.

$P(x_j^i|c)$ is the probability of the word j in the class c.

The prior probability $P(c)$ of a class can be calculated as follow:

$$P(c) = \frac{N_c}{N}$$

Where:

N_c is the number of reviews in the class c.

N is the total of all movie review.

Then $P(x_j^i|c)$ is the probability of the word j in the class c can be calculated as follow:

$$P(x_j^i|c) = \frac{N_{jc} + 1}{N_c + |vocabulary|}$$

Where:

N_{jc} is the number of occurrences of the word j in a class c.

N_c is the number of occurrences of all words in a class c.

$|vocabulary|$ is the total number of all unique words in the dataset.

Finally, when classifying a new document, whether the document is classified as a class c can be done using the formula below:

$$P = argmaxP(c) \prod_{j=1}^{m} P(x_j^i|c)$$

Then highest value of probability is selected as the class of the reviews. The limitation is that all the features on MNB model are assumed to be independent. Although it might sound great in theory. However, in real life, it hardly finds a set of independent features. The MNB algorithm is used by calling MultinomialNB() method of ”sklearn”.

3.8 Evaluation

Performance evaluation is a very inevitable step to find out the performance of machine learning models. The confusion matrix measures the performance, which is a 2x2 sized matrix. The confusion matrix is used to calculate the values of accuracy, recall, precision and F1-score. For binary classification, the confusion matrix is shown on Table 1:

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Table 1: Confusion Matrix

where :

True Positive (TP) is the total of positive samples that are accurately predicted.

True Negative (TN) is the total of negative samples that are accurately predicted.

False Positive (FP) is the total of positive samples that are inaccurately predicted.

False Negative (FN) is the total of negative samples that are inaccurately predicted.

Accuracy is a common metric to evaluate the performance of classification models. It describes the probability that a machine learning algorithm predicts data correctly. Accuracy is depicted by the ratio of the number of correctly classified data in the total number of data [19].

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

The recall is described by the number of actual positive events with respect to the number of data that got predicted as positive. Given a confusion matrix, recall may be calculated by the TP value with respect to the sum of TP and FN.

$$recall = \frac{TP}{FN + TP}$$

Precision is one of popular metrics that is used to evaluate information retrieval, classification, and pattern recognition-related problems. It is depicted by the proportion of relevant observations in regard to retrieved observations. Given a confusion matrix, precision is calculated by TP with respect to the total number of TP and FN.

$$precision = \frac{TP}{FP + TP}$$

F-1 score is a metric which combines precision and recall is given by

$$F - 1score = \frac{recall \times precision \times 2}{recall + precision}$$

These values above are between 0 and 1. The higher these values are, the better the performance of the classification model is.

In this paper, there are five classes rather than two classes. The formulas for the performance evaluation in multi-class classification are as same as the formulas in binary classification. However, the difference is that, in multi-class classification problem, the accuracy, recall, precision, and F-1 score in each class are calculated. Then use weighted-average method to calculate the final values. The weighted-average method is to sum the values of all classes after multiplying their respective class weights. The weighted-average method considers unbalanced distribution, and its value is mainly influenced by the majority class.

4 Result analysis

4.1 Data Analysis

The dataset is a DouBan movie review dataset. The dataset contains 520,000 reviews from 9,609 movies in total. After the data preprocessing, 318,180 movie reviews are still kept. Figure 4 shows the distribution of the rating on datasets. This data is structured in EXCEL files, including columns such as “COMMENT ID” , “USER NAME” , “MOVIE ID” , “CONTENT” , “VOTES” , “COMMENT TIME” and “RATING” . The ”CONTENT” on dataset shows the Chinese movie reviews, while the ”RATING” on dataset is the ratings of corresponding movie reviews. However, in this paper, only columns “CONTENT” and “RATING” are considered and are taken as input features to define the supervised learning problem, because sentiment analysis is viewed as a supervised learning task. The DouBan rating system allows viewers to rate the movies on a scale from 1 to 5. The are five classes from small to large represent extremely negative, negative, neutral, positive and extremely positive ones, respectively. If the rating is 1, the viewer does not satisfy the movie. If the rating is 5, people think the movie is extremely good. The higher rating is, the better review is.

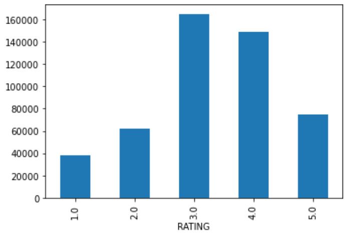

Figure 5: Data Distribution

	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Number of reviews	37,829	74,522	165,052	149,128	62,383

Table 2: The number of reviews

The number of movie reviews for each rating is shown on Table 2. There are most reviews on rating 3 and rating 4, 165,052 and 149,128, respectively. The number of movie reviews on rating 2 and rating 5 are similar, 74,522 and 62383, respectively. The number of rating 1 is the fewest, which is 37,829. It is an unbalanced distribution. In order to construct the machine learning models, this dataset is splited into 80:20 training and testing data. That means 80% of the dataset is used for training purposes, and 20% of the dataset is used for testing purposes. The set of training data is used to train the considered classifier, while test data is used to evaluate the model performance, such as accuracy [18].

4.2 Performance Analysis

After running the five machine learning algorithms successfully on the DouBan movie review dataset, the results are shown on Table 3. This Table 3 shows the summary of Accuracy, Precision, Recall, and F1-Score matrices.

Evaluation Measurement	Accuracy	Precision	Recall	F1-score
Support Vector Machine	42.83%	45.02%	42.84%	38.59%
K-Nearest Neighbor	36.03%	37.26%	36.03%	26.22%
Logistic Regression	42.98%	43.31%	42.99%	40.16%
Multinomial Naive Bayes	60.37%	40.42%	41.14%	38.47%
Random Forest	41.76%	43.99%	41.76%	36.28%

Table 3: Comparison of Results for Different Classifiers

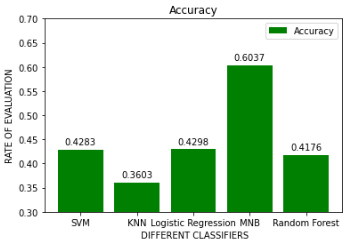

Figure 6: Accuracy of Different Classifiers

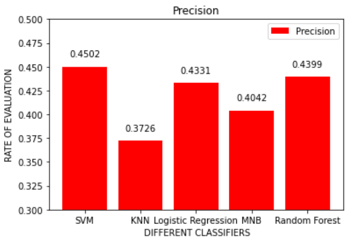

Figure 7: Precision of Different Classifiers

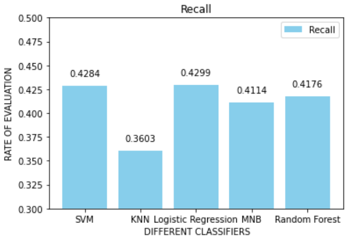

Figure 8: Recall of Different Classifier

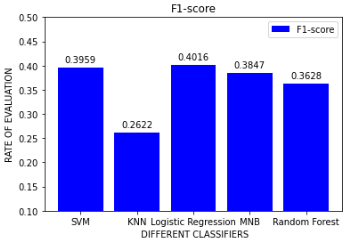

Figure 9: F1-score of Different Classifier

The best accuracy is achieved when TF-IDF is used and Multinomial Naive Bayes (MNB) is chosen as the classifier. The highest accuracy is 60.37%, which means that the MNB method has correctly classified 313,924 data. The correctly classified data is divided by 520,000, resulting in an accuracy of 60.37%. The accuracy of the MNB method with TF-IDF is best among those 5 models. It is because that the MNB method can work better than others when training with a considerably complex datasets, such as the DouBan movie reviews dataset. In this experiment, there are 163,887 features and 520,000 movie reviews, which means that this dataset is complex. The precision value is 40.42%, which means that the number of relevant observations with respect to retrieved observations is 40.42%. Moreover, the recall value of MNB is 41.14%, which means 41.14% of movie reviews on Douban are correctly identified as having the condition. The value of recall on MNB is higher than the value of precision. That means, for each class, positive samples and a portion of negative samples are predicted as positive samples, where the positive sample is the sample in the given class and other classes are defined as negative samples. F-1 score is 38.47%. F-1 score is the mean of the values of recall and precision, so it is an important parameter when comparing MNB with the other four models. The higher F1-score is, the more robust model is.

Additionally, the accuracy, precision, recall, and F-1 score values of the Logistic Regression model are 42.98%, 43.31%, 42.99%, and 40.16%, respectively. The accuracy of the Logistic Regression model is lower than that of the MNB model but higher than the other three models. Moreover, the Logistic Regression model gets the highest recall and F1-score values among those five models. Thus, the Logistic Regression model is the most robust among those five models. It is because the coefficients of Logistic Regression can infer the importance of each feature and the associated direction. Moreover, Logistic Regression outputs well-calibrated probabilities along with classification results, which improves the F-1 score of the prediction as well.

As for the Support Vector Machine(SVM) model, accuracy, precision, recall, and F-1 score value are 42.83%, 45.02%, 42.84%, and 38.59%, respectively. The SVM model has the highest precision value among those five models. In general, models with higher precision value can better distinguish negative samples from positive samples. Thus, compared with the other 4 models, the SVM can distinguish negative samples from positive samples greatly.

For the Random Forest model, the accuracy, precision, recall, and F-1 score value are 31.76%, 43.99%, 41.76%, and 36.28%, respectively. The all the values in the random forest are lower than the values in SVM, because the linear kernel of SVM is very suitable for text classification problems, such as sentiment analysis on movie reviews. Hence, SVM classifier can classify text data with relatively high accuracy result. Although the values of Random Forest model are lower than values of SVM, Logistic Regression and MNB model, the results in Random Forest are higher than results of K-Nearest Neighbor (KNN) model. It is because Random Forest can automatically deal with noise, missing values and outliers. In general, Random Forest is usually robust to outliers and less influenced by noise, while KNN model are sensitive to the outliers and noise.

Moreover, K-Nearest Neighbor (KNN) gets the worst results among the five models, where the accuracy, precision, recall, and F1-score value are 36.03%, 37.26%. 36.03% and 26.22%, respectively. In particular, KNN model gets the lowest precision among the 5 models, which indicates that predicted result in KNN model is very far from the truth.

Hence, the MNB method produces the best accuracy with results above 60%, which is far higher than other models. Moreover, the Logistic Regression model is the most robust model among those 5 models, while KNN model has the worst performance.

From table 3, it is obvious that the accuracy of different classification models are between 36.03% to 60.37%. These data in the dataset are trained and classified into 5 classes. Thus, all of the machine learning models work, because all the accuracy of these models are higher than 0.2, where 0.2 is the classical probability of prediction of the correct class. However, the number of classes for classification still affect the accuracy of these models. Compared with some previous sentiment analyses of movie reviews with binary classification, the accuracy in this paper are relatively low. It is because more classes for categorization decrease the probability of mapping a review into each class [19]. The reviews in the dataset are classified into five classes, which contributes to the relatively poor accuracy of results for sentiment analysis of the movie reviews.

Moreover, some limitations on vectorization affect the accuracy of these models as well. Firstly, the huge amount of data in the dataset extract a lot of features from text into classifiers as input. In this experiment, 22,234 words are extracted as features. For each sentence, there are no more than 45 words, so the matrix of reviews is sparse. The sparse matrix is a matrix where most of the elements are zero. The sparse matrix seriously affects the memory and computing resources, causes errors when training the models and decreases the accuracy of models. Secondly, the limitations of the TF-IDF method decreases the accuracy. For example, the features extracted by the TF-IDF method only depend on the frequency of the words in sentences. Besides, some unfamiliar words are selected as features, so that the matrix becomes sparser and then the accuracy is decreasing. The third limitation is that the TF-IDF method extracts the words which are independent of other words, and ignores the order and grammar of the words, so that models cannot understand the meaning of sentences accurately.

Besides, some shortcomings on pre-processing have impact on the performance of sentiment analysis. Firstly, the classification model's performance is influenced by the stopwords list, especially in the sentiment analysis of Chinese reviews. In this paper, the list for stopwords derived from HIT, which is recognized as the best stopwords list for Chinese text but only includes the basic and limited Chinese stop words. Thus, only limited stopwords are removed before training models, so the accuracy is reduced. Secondly, in word segmentation process, although Jieba is recognized as the best Python Chinese word segmentation module, it also creates lots of unfamiliar words or wrong phrases. Thus, it is not surprising that there have been some misclassifications. For example, in this experiment, Jieba divides the reviews and creates the word “很不” which means “extremely not” in English. However, there are no such words in Chinese speaking and writing. For those mistaken Chinese words, it increases the features and complexity when training the models and decreases the result values. Moreover, “extremely not” has the same meaning as “not” in Chinese, but these two words become two features when training models. Thus this limitation increases the complexity and decreases the efficiency and accuracy. Thirdly, classification could not detect sarcasm, which leads to errors when training the classification models.

Additionally, the performance of models in this experiment are affected by limitations of practice. Firstly, the data distribution of ratings in the dataset is unbalanced, which affects the results of classification. As the above Figure 5, most of the movie reviews are in ratings 3 and 4, which leads to unbalanced results. The results are mainly influenced by majority classes. Secondly, the human factor also affects the results. Different people have different personalities and different criteria when rating the movies. For example, some people rate 5 stars when they like this movie, but others only rate 4 stars, because they are strict about the movie and think the movie could be better. Machine learning models are hard to imitate the thinking of human.

5 Conclusion

With rapidly increasing a huge amount of movie reviews on DouBan website, sentiment analysis of the movie reviews on this website is important. Through the sentiment analysis, public opinions about the movies can be foretasted and ratings of movie can be predicted. This research article focuses on predicting movie ratings based on movie reviews with the sentiment analysis method. The proposed methodology has used the Douban dataset, where Douban is the one of most popular movie review aggregation sites for Chinese to write down their reviews after watching movies and is similar to IMDB. The classification process is demonstrated by classifying the reviews in the dataset into 5 classes, which are the ratings of the movie and on the scale of 1 to 5. Those 5 classes from 1 to 5 represent extremely negative, negative, neutral, positive, and extremely positive ones, respectively. Pre-processing of the dataset is inevitable before feeding reviews to the classifier model. For example, Stopwords should be removed. TF-IDF method is chosen for text representation in the work. After that, the sentiment analysis on the Douban movie reviews dataset is performed by applying five machine learning models. Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Multinomial Naive Bayes, and Random Forest are used in this experiment. These results are then compared based on different evaluation metrics. The Multinomial Naive Bayes classifier achieves the highest classification accuracy of 60.37%. Logistic Regression has the largest F1-score (40.16%), which means the Logistic Regression model is the most robust model among those five models. K-Nearest Neighbor model gets the worst result, so K-Nearest Neighbor is not suitable for multi-class classification problem.

In the future, several improvements could be done to improve the performance. Firstly, research directions may be in extended experiments involving word embeddings like Word2vec. It is because that Word2vec improves the inadequacies of the TF-IDF. Word2Vec captures the semantic similarity between the words and words with similar meanings are placed in close proximity in the vector space [4]. Secondly, introducing ensemble classifiers or deep learning algorithms to sentiment analysis of DouBan movie reviews to improve classification accuracy. For example, LSTM, CNN and CNN-LSTM could be employed to solve this multi-class classification problem. Thirdly, improve the accuracy of creation of tokens on JIEBA, so that there are fewer unfamiliar words and wrong phases as features when training the models. Fourthly, introducing more lexicon features to the feature subset can increase the classification accuracy.

References

[1] Warda Ruheen Bristi, Zakia Zaman, and Nishat Sultana. Predicting imdb rating of movies by machine learning techniques. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Computing, Communication and Networking Technologies (ICCCNT), 2019 10th International Conference on*, pages 1 – 5, 2019.

[2] Ramadhan Nur Ghaniaviyanto and Ramadhan Teguh Ikhlas. Analysis sentiment based on imdb aspects from movie reviews using svm. *Sinkron*, 7(1):39 – 45, 2022.

[3] Kamil Topal and Gultekin Ozsoyoglu. Movie review analysis: Emotion analysis of imdb movie reviews. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 1170 – 1176, 2016.

[4] Sandesh Tripathi, Ritu Mehrotra, Vidushi Bansal, and Shweta Upadhyay. Analyzing sentiment using imdb dataset. *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Computational Intelligence and Communication Networks (CICN), 2020 12th International Conference on*, pages 30 – 33, 2020.

[5] Saikiran Gogineni and Anjusha Pimpalshende. Predicting imdb movie rating using deep learning. *2020 5th International Conference on Communication and Electronics Systems (ICCES), Communication and Electronics Systems (ICCES), 2020 5th International Conference on*, pages 1139 – 1144, 2020.

[6] Md. Rakibul Haque, Salma Akter Lima, and Sadia Zaman Mishu. Performance analysis of different neural networks for sentiment analysis on imdb movie reviews. *2019 3rd International Conference on Electrical, Computer Telecommunication Engineering (ICECTE), Electrical, Computer Telecommunication Engineering (ICECTE), 2019 3rd International Conference on*, pages 161 – 164, 2019.

[7] Savitha Mathapati, Amulya K Adur, R Tanuja, S H Manjula, and K R Venugopal. Collaborative deep learning techniques for sentiment analysis on imdb dataset. *2018 Tenth International Conference on Advanced Computing (ICoAC), Advanced Computing (ICoAC), 2018 Tenth International Conference on*, pages 361 – 366, 2018.

[8] Velery Virgina Putri Wibowo, Zuherman Rustam, and Jacub Pandelaki. Classification of brain tumor using k-nearest neighbor-genetic algorithm and support vector machine-genetic algorithm methods. *2021 International Conference on Decision Aid Sciences and Application (DASA), Decision Aid Sciences and Application (DASA), 2021 International Conference on*, pages 1077 – 1081, 2021.

[9] H. S. and R. Ramathmika. Sentiment analysis of yelp reviews by machine learning. *2019 International Conference on Intelligent Computing and Control Systems (ICCS), Intelligent Computing and Control Systems (ICCS), 2019 International Conference on*, pages 700 – 704, 2019.

[10] Sudha Shanker Prasad, Jitendra Kumar, Dinesh Kumar Prabhakar, and Sukomal Pal. *Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree.*, volume 9468 of *Lecture Notes in Computer Science. 9468*. Springer International Publishing, 2015.

[11] Bhanu Prakash Reddy Guda, Mashrin Srivastava, and Deep Karkhanis. Sentiment analysis: Predicting yelp scores. 2022.

[12] Asriyanti Indah Pratiwi and Adiwijaya. On the feature selection and classification based on information gain for document sentiment analysis. *Applied Computational Intelligence and Soft Computing*, 2018, 2018.

[13] Atiqur Rahman and Md. Sharif Hossen. Sentiment analysis on movie review data using machine learning approach. *2019 International Conference on Bangla Speech and Language Processing (ICBSLP), Bangla Speech and Language Processing (ICBSLP), 2019 International Conference on*, pages 1 – 4, 2019.

[14] Mais Yasen and Sara Tedmori. Movies reviews sentiment analysis and classification. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Electrical Engineering and Information Technology (JEEIT), 2019 IEEE Jordan International Joint Conference on*, pages 860 – 865, 2019.

[15] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117 – 126, 2016.

[16] Vivek Narayanan, Ishan Arora, and Arjun Bhatia. Fast and accurate sentiment classification using an enhanced naive bayes model. 2013.

[17] Sakshi Bansal, Chetna Gupta, and Anuja Arora. User tweets based genre prediction and movie recommendation using lsi and svd. *2016 Ninth International Conference on Contemporary Computing (IC3), Contemporary Computing (IC3), 2016 Ninth International Conference on*, pages 1 – 6, 2016.

[18] Saeed Mian Qaisar. Sentiment analysis of imdb movie reviews using long short-term memory. *2020 2nd International Conference on Computer and Information Sciences (ICCIS), Computer and Information Sciences (ICCIS), 2020 2nd International Conference on*, pages 1 – 4, 2020.

[19] Nisha Rathee, Nikita Joshi, and Jaspreet Kaur. Sentiment analysis using machine learning techniques on python. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Intelligent Computing and Control Systems (ICICCS), 2018 Second International Conference on*, pages 779 – 785, 2018.