Sentimental Analysis Applied on Movie Reviews

SichangSu*

Department of Aeronautics and Astronautics, Zhejiang University, Zhejiang, China

*Corresponding author: 3190102891@zju.edu.cn

Abstract. Nowadays, Natural Language Processing has received the widespread attention from the natural sciences, and sentimental analysis is one of the most widely used NLP applications. In the age of big data, how to find the required information accurately and quickly has become the hotspot of current research. Based on the movie reviews of two movies from the same series, this paper studies the sentimental trend of movies reviews, in order to help the audience obtain a reference for movie choices. Term frequency-Inverse Document Frequency (TF-IDF) algorithm is applied to evaluate the importance of words in the reviews, and TextBlob sentiment analysis library of Python software is used to grade the sentiment scores of the two films. Finally, the sentiment score graph is drawn, which provides a strong support for the further identification of the movie characteristics of two films from the same series. What's more, Support vector machines (SVM) model is utilized to do the classification of the movie reviews and achieved 85.2% accuracy.

Keywords: Sentimental Analysis, IMDB Movie Review, Sentimental Dictionary, TextBlob, SVM Model.

1. Introduction

Natural language processing (NLP) is the conversion of spoken and written human language into computer-understandable machine language. It is a model and algorithmic framework for researching the power of language, as well as a field that combines computer science and linguistics.

One of the most widely used NLP applications is sentiment analysis, which is also known as dispositional analysis, sentiment mining, subjectivity analysis, and etc. It is the process of deciphering, understanding, extrapolating from, and making sense of the subjective and emotionally laden texts [1].

There are two main groups of sentiment analysis techniques for short reviews: sentiment dictionary matching based method and machine sentiment analysis based on Python statements. First, sentiment analysis based on sentiment dictionary matching refers to splitting and deactivating the text content of the movie review, and then it can use Python software combined with sentiment dictionary to match sentiment words to find the positive and negative words. Second, machine sentiment analysis based on ML techniques mainly uses software to filter the statements with the positive and negative sentiment in the text, and then uses machine learning method to analyze the filtered statements [2].

In current scenario, Natural Language Processing has received the widespread attention from the natural sciences, represented by computer science, to the social sciences. To be more specific, this technology has demonstrated its indispensable value in issues such as news communication, opinion management, and opinion analysis [3].

Everyone is a creator and user of information while more and more companies are trying to mine the valuable information from data to solve business problems [4]. In the era of big data, how to find the required information accurately and quickly has become the hotspot of current research. This article is based on this specific intention to conduct sentimental analysis from the different angles for the same series of film reviews, in order to help the audience obtain a reference for movie choices. To demonstrate the effectiveness of the utilized method, two movies with the high scores on IMDB, The Lord of the Rings: The Return of the King and The Lord of the Rings: The Fellowship of the Ring, were selected for the article.

In the proposed work, TextBlob is applied to do sentiment analysis of movie reviews. Besides, the work of high-frequency word extraction of text has been carried out too, in order to estimate the sentimental trend of movies.

The following arrangement characterizes the remaining sections: The related publications that describe the use of sentiment analysis to analyze and categorize movie reviews are covered in Section II. The proposed methodology, which comprises the TF-IDF algorithm, sentiment dictionary, and SVM model, is covered in Section III. The experiment's results are presented in Section IV. Section V then explains the findings and offers a judgment.

2. Related work

Nowadays, Machine Learning and Deep Learning methods are predominantly applied to classify the sentiments and conduct sentimental analysis [5]. The numerous researchers are concentrating on sentimental analysis for online product reviews including movie reviews.

Huy Tien Nguyen, et al. [6] discovered the issue that merging these benefits of CNN and LSTM model classifying sentiments into one model was difficult owing to over-fitting in training and eventually established a freezing strategy to extract opinion vectors from LSTM and CNN on five distinct datasets, including the IMDB huge movie reviews dataset, and attained over 93 percent accuracy. Zeeshani Shaukat, et al. [7] successfully completed the task of opinion mining from movie reviews using neural networks trained on the Stanford Movie Review Database together with two long series of positive and negative phrases, and they did it with an accuracy rate of 91.9 percent. By analyzing and contrasting five Machine Learning and Deep Learning algorithms using the specific sentiments from film reviews, Onalaja Samuel, et al. [8] came to the conclusion that giving more weight to particular characteristics and genres results in the five models' accuracy increasing [8]. B. Selvakumar, et al [5] .'s classification of the sentiments was more precise and effective because they took into account the contextual relationship across the full sequence of words. To analyze the IMDB movie reviews and the Amazon fine food reviews, they utilized a multi-self-attention-based BERT model and achieved 94 percent accuracy. According to Keerthi Kumar et al. [9] , the usage of hybrid features, which are produced by combining machine learning features (TF, TF-IDF) with lexical features (Connotation, Positive-Negative word count), outperforms standard classifiers like KNN, SVM, Nave Bayes and Maximum Entropy according to complexity and accuracy.

3. Methodology

This project aims to do a more thorough analysis of movie reviews' emotions. SVM model is employed in this article to categorize the sentiments. The online movie reviews from IMDB are used for experimentation in the proposed sentiment classification task. The proposed system's architecture is shown in Figure 1. The following list outlines the steps for the sentiment analysis:

- Data Collection
- Data Pre-processing
- Sentiment Analysis using TextBlob
- SVM Model Training

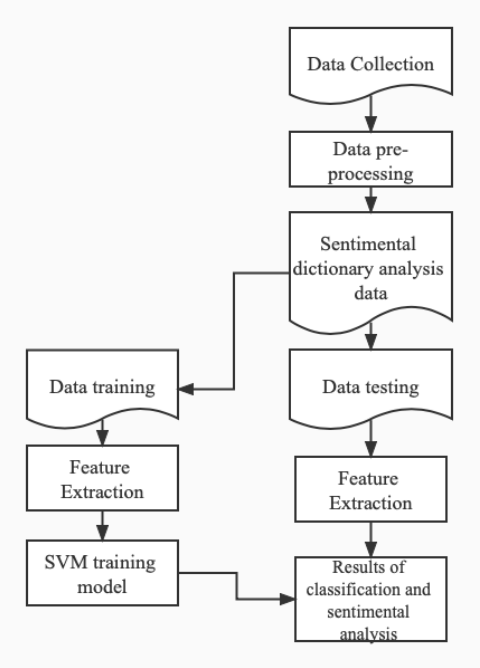

Figure 1. Overall flow of system

3.1. Data collection

In this study, Crawler tool Pyspider is utilized to crawl the movie reviews from IMDB of The Lord of the Rings: The Return of the King and The Lord of the Rings: The Fellowship of the Ring. After de-duplication and filtering, 1,500 reviews are left as experimental data for each movie.

3.2. Data pre-processing

3.2.1 Word segmentation using sentiment dictionary

HowNet sentiment dictionary is selected as the basic sentiment dictionary. The steps of using the sentiment dictionary for word segmentation are as follows:

- The positive sentiment words and positive evaluation words in HowNet are combined, and imported into the Python software as a positive sentiment dictionary for backup.
- The negative sentiment words and negative evaluation words in HowNet were combined, and the sense dictionary is imported into the Python software for use.

The degree words can be classified through HowNet sentiment dictionary. For example, the meaning words of the degree words ‘extremely’ and ‘most’ are divided into the first category. The meaning words of ‘very’ are divided into the second category. The meaning words of ‘comparative’ are divided into the third category. The meaning of the words is divided into the fourth category, and imported into the Python as a degree-sentiment dictionary. The Python wordninja library is utilized for word segmentation.

3.2.2 Removing deactivated words

The deactivation word processing is performed on the documents after the movie reviews are divided into words. In this paper, stop words list from NLTK is chose to remove deactivated words for the movie reviews.

3.2.3 Text feature extraction

Word frequency analysis is carried out for the review data of the two movies, and the high frequency words of the two movies are counted.

3.2.4 Term frequency-Inverse Document Frequency algorithm

Term frequency-Inverse Document Frequency (TF-IDF) algorithm reflects the internal features of a text through term frequency, which can be used to evaluate the significance of a word to a text. It is actually a combination of term frequency and inverse document frequency. The formula of this algorithm is presented as follows [2]:

$$TF_\omega \cdot IDF_\omega = \frac{a}{b} \cdot \log \frac{N}{n_\omega + 1} \tag{1}$$

where ω stands for the word or words calculated, a stands for the number of occurrences of ω feature items in the text, b stands for the number of occurrences of all feature items in the text, N stands for the total number of documents in the corpus, and n_ω stands for the number of documents containing ω in the corpus.

The relevant parameters from the dataset would be substituted into the equation, with the purpose of getting the film review TF - IDF value table.

3.3. Sentiment analysis using TextBlob

TextBlob library in Python is based on a sentiment dictionary, which can be utilized to perform various natural language processing (NLP) tasks, such as sentiment analysis, lexical annotation, text translation, noun extraction, and etc. It classifies the text to be analyzed into two categories to be processed. Emotions are generally positive, neutral and negative. However, in this paper, only the positive and negative sentiments of the text are taken into consideration. In TextBlob sentiment analysis, the range of sentiment score is [-1, 1]. The closer to 1 the sentimental score is, the more positive the sentiment is. Meanwhile, the closer to -1 the sentimental score is, the more negative the sentiment is. The operation step of this sentimental analysis is to calculate the sentiment score of each movie review and generate a sentiment score bar chart.

3.4. SVM model training

3.4.1 Feature selection

The key of machine learning-based sentiment analysis is feature selection, which is related to the accuracy of sentiment classification. The common feature choices are: unigram features, bigram features, trigram features, word frequency, lexicality, and sentiment words [10]. The feature dimensions of unigram, bigram and trigram features have relation to the volume of the corpus. When the corpus is large, the feature dimensions will reach the thousand-dimensional level, which is difficult to handle. Word frequency can reflect the importance of a word, but not all words are related to text sentiment. Therefore, sometimes the introduction of word frequency will lead to inaccuracy in the data. In this paper, five feature dimensions are chosen: wordiness, degree adverbs, negation, positive emotion words and negative emotion words, in which a text is composed of multiple words and their wordiness and wordiness plays a big role. Sentimental words are the key core of a text's sentimental classification, and negation words usually reverse the sentimental polarity of a text. Meanwhile, degree adverbs can also change the intensity of emotion words when both positive and negative sentiment words appear in a text. it is hard to establish the sentiment tendency of the text if only the polarity of the sentiment word is relied on. However, degree adverbs are able to help to make a choice [11].

When selecting text features, for each dimension the specific meaning is shown in Table.1.

Table 1. Dimension characteristic meanings

Wordiness	Meaning
Wordiness	Counting the number of adjectives, nouns, verbs and exclamations
Degree adverbs	Counting the cumulative weight of degree adverbs
Negative words	Counting the number of negative words
Positive emotional words	Counting the number of positive emotional words
Negative emotional words	Counting the number of negative emotional words

3.4.2 SVM model

Support vector machines (SVM) is a new classification method developed in recent years, which mainly solves text classification problems [12] . The method uses a supervised learning approach to model the binary classification problem. When linearly separable, the samples are separated by finding an optimal hyperplane. In contrast, when linearly non-separable, they are transformed into linearly separable using kernel functions, usually Sigmoid kernel, radial basis function kernel, polynomial kernel and Laplace kernel. The SVM problem is converted into a convex quadratic programming problem using the maximum interval hyperplane, which is the ideal hyperplane where the distance to the closest data point is maximum on both sides.

The support vector machine model's fundamental premise is to determine the largest geometric interval between two types of samples. The effect and generalizability of the classifier are enhanced by minimizing the upper bound of the algorithm's error by maximizing the geometric interval [13].

4. Results

In this work, 3,000 IMDB movie reviews of two movie is chosen for the experiment. HowNet sentimental dictionary is used to do word segmentation. After removing deactivated words, word frequency of the movie reviews of the two movies is analyzed. This study obtained the conclusion that "movie" and "ring" as the first and second high frequency words for The Lord of the Rings 1 while "movie" and "Jackson" as the first and second high frequency words for The Lord of the Rings 3. The top 10 high frequency words and the number of occurrences are shown in Table.2. Besides, the relevant data of movie reviews are substituted into Equation (1), the film reviews TF - IDF values are presented in Table.3. What's more, the sentiment scores of movie reviews are attained through the TextBlob library of Python, in order to draw sentiment score maps for two movies, which clearly shows whether the two movies tend to be positive or negative. Figure 2 presents the the sentiment scores for The Lord of the Rings1, while Figure 3 presents the sentiment scores for The Lord of the Rings3. Finally, SVM model is utilized to classify sentiments after selecting word features. The trials are carried out using a 70:30 train-to-test split ratio. Figure 4 shows the result of SVM model for the movie reviews.

Table 2. Top 10 high frequency words in movie reviews

The Lord of the Rings1		*The Lord of the Rings3*	
Words	Frequency	words	Frequency
movie	6469	movie	7349
ring	2392	Jackson	1383
Jackson	1188	trilogy	1284
book	1155	king	1181
story	1117	Frodo	1117
time	1035	time	1035
lord	990	story	983
fellowship	941	return	969
characters	913	battle	909
Frodo	855	lord	870

Table 3. Movie Review TF - IDF Value (top 10 high frequency words)

The Lord of the Rings1				*The Lord of the Rings3*			
Words	TF	IDF	TF-IDF	Words	TF	IDF	TF-IDF
movie	0.0235	0.3074	0.0072	movie	0.0266	0.3434	0.0092
ring	0.0087	0.3945	0.0034	Jackson	0.0050	0.7911	0.0040
Jackson	0.0043	0.8045	0.0035	trilogy	0.0047	0.7202	0.0034
book	0.0042	0.6475	0.0027	king	0.0043	0.5379	0.0023
story	0.0040	0.7161	0.0029	Frodo	0.0040	1.0788	0.0044
time	0.0038	0.6286	0.0024	time	0.0038	0.6274	0.0024
lord	0.0036	0.9096	0.0033	story	0.0036	0.7664	0.0027
fellowship	0.0034	1.0366	0.0035	return	0.0035	0.9979	0.0035
characters	0.0033	1.0217	0.0034	battle	0.0033	0.9399	0.0031
Frodo	0.0031	1.2730	0.0039	lord	0.0032	1.0441	0.0033

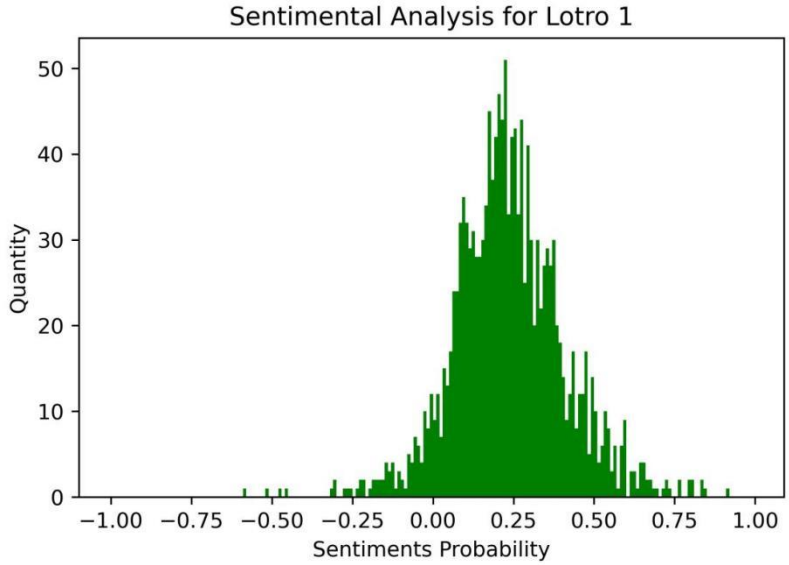

Figure 2. Sentiment Score Map for *The Lord of the Rings1*

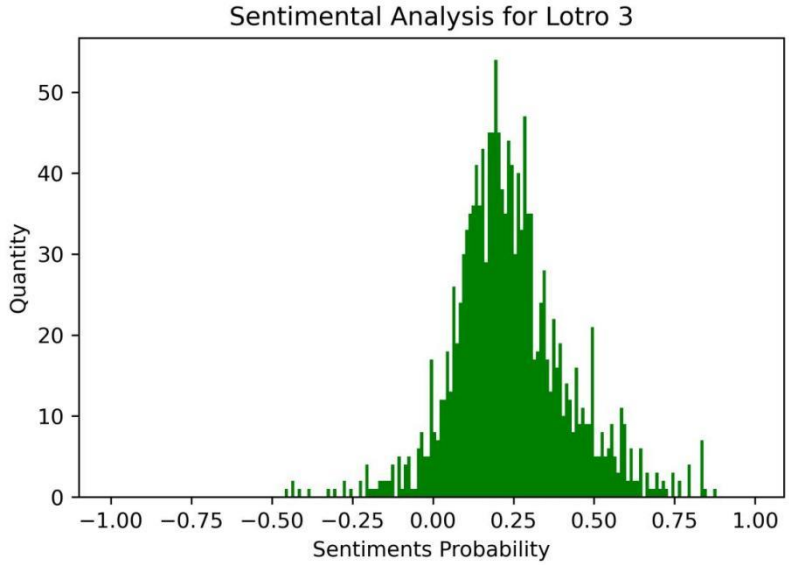

Figure 3. Sentiment Score Map for *The Lord of the Rings3*

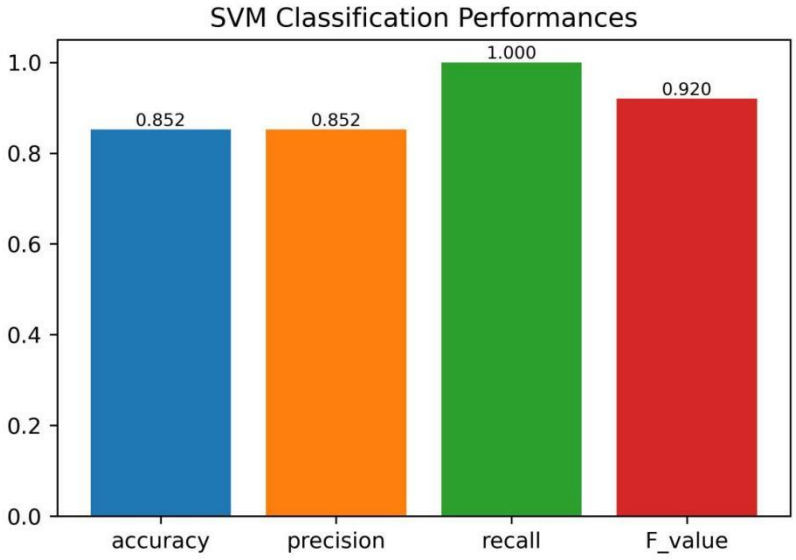

Figure 4. SVM Classification Results

5. Conclusion

The larger the TF-IDF value, the stronger the importance of the word in the text, which could help researchers to evaluate the importance of words in reviews. For the sentimental score map, it takes the abscissa 0 as the boundary between the positive and negative evaluation values of movie reviews. The sentimental score distribution in the [-1, 0) interval is negative evaluation, the sentimental score distribution in (0, 1] is positive evaluation, and the ordinate indicates the amount of evaluation. The sentimental score of the movie The Lord of the Rings1 is mainly concentrated in the interval [0.15, 0.25], and the emotional score of the movie The Lord of the Rings3 is obviously concentrated in the interval [0.1, 0.3]. This reflects the The Lord of the Rings3 has a little stronger positive emotions than The Lord of the Rings1, which reflects why The Lord of the Rings3 ranks higher than The Lord of the Rings1 in IMDB. The SVM model gives accuracy of 85.2%, which proves that doing words feature selection before training the SVM model would obtain a precise results in sentimental classification. Future work on this model can include experimenting with various sentimental analysis algorithms for movie reviews, with the intention of higher accuracy and efficiency. In this paper, only 1,500 front-row movie reviews for each movie were crawled from IMDB website, so there is a lack of volume and single data in movie review data crawling. The research area and angle can be further explored in the later research.

References

[1] ZhaoYa'ou,ZhangJiachong,LiYibin,FuXianrui,Sheng Wei. Sentiment analysis using embedding from language model and multi-scale convolutional neural network[J]. Journal of Computer Applications,2020, 40(3):651-657.

[2] BaoShuhua, ShiYingxin. Sentiment Analysis of Domestic Movie Reviews Under the Condition of Big Data[J]. Journal of Hulunbuir University,2022, 30(2):126-131.

[3] WuXiaokun, ZhaoTianfang. Application of Natural Language Processing in Social Communication:A Review and Future Perspectives[J]. Computer Science, 2020, 47(6):184-193.

[4] ShaoXiaoqing. Data Mining and Analysis Based on Python Movie Review[J]. Information Reconding Material,2021, 22(10):224-226.

[5] B. Selvakumar, B. Lakshmanan, Sentimental analysis on user's reviews using BERT[J]. Materials Today: Proceedings, 2022, 62(7):4931-4935.

[6] H.T. Nguyen, M.L. Nguyen, An ensemble method with sentiment features and clustering support, Neurocomputing,2019, 370:155–165, https://doi.org/10.1016/j.neucom.2019.08.071.

[7] Z. Shaukat, A.A. Zulfiqar, C. Xiao, M. Azeem, T. Mahmood, Sentiment analysis on IMDB using lexicon and neural networks, SN Appl Sci, 2020,2, https://doi.org/10.1007/s42452-019-1926-x.

[8] O. Samuel, R. Eric, Y. Bosang, Aspect-based Sentiment Analysis of Movie Reviews, SMU Data Science Review,2021, 5 (3) ,Article 10.

[9] K. Kumar, B.S. Harish, H.K. Darshan, Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method, IJIMAI 5, 2019, 109, https://doi.org/10.9781/ijimai.2018.12.005.

[10] Li Bin, Yang Qiang, XueXiangyang. Transfer learning for collaborative filtering via a rating-matrix generative model [C]/ / Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 617 – 624.

[11] Fan Zhen, Guo Yi, Zhang Zhenhao, Han Meiqi. Sentiment analysis of movie reviews based on dictionary and weak tagging information [J]. Journal of Computer Applications, 2018, 11: 3084-3088.

[12] MirbakhshNima, Ling Charles X. Improving top-n recommendation forcold-start users via cross-domain information [J]. ACM Transactions on Knowledge Discovery from Data. 2015. 9 (4): Article No.33.

[13] Wang Wentao, Zhang Shibao. Emotion Analysis of Micro-blog Netizens Based on Emotion Dictio nary and SVM. Modern Information Technology, 2021, 24-27+31. doi:10.19850/j.cnki.2096-4706.2021.24.007.