Text Sentiment Analysis for Movie Theme Forecasting Based on Natural Language Processing

Xuhui Song

School of Mathematics and Statistics. Henan University of Technology. Zhengzhou. Henan. China

songxuhui1103@outlook.com

Abstract. With the advancement of the internet, online film reviews have emerged as a significant platform for audiences to share their viewing experiences. This study, leveraging short review data from Douban movies, proposes a film genre prediction method that integrates text sentiment analysis and high-frequency word matching. By scraping short reviews from Douban's Top 250 films and utilizing Python's jieba package for word segmentation, high-frequency terms were extracted and matched with genres using a constructed sentiment dictionary. Experimental results demonstrate that this approach effectively identifies mainstream genres such as drama and romance, though its performance in detecting niche genres requires further enhancement. The key innovation lies in directly extracting genre-related features from audience reviews and improving keyword recognition accuracy through sentiment analysis. While the method proves effective to a certain extent in predicting film genres, there remains room for improvement in both accuracy and generalizability. This study not only introduces a novel technical framework for film genre analysis but also provides empirical insights into the relationship between audience genre preferences and emotional responses. Future research could explore the integration of deep learning models to further enhance the precision and robustness of genre prediction.

Keywords: Python, Web Crawling, Movie Themes, Douban, NLP.

1. Introduction

With the rapid development of the Internet and social media, online movie reviews have become an important channel for audiences to express their viewing experiences and emotional tendencies. Douban, as one of the most influential movie rating and review platforms in China, has accumulated a vast amount of user-generated content, which provides rich research materials for movie theme analysis and sentiment mining [1, 2]. However, most of the current research based on Douban movie short reviews focuses on rating prediction or sentiment polarity analysis, and few studies have explored how to use text sentiment analysis technology to predict and classify movie theme types.

This paper uses the method of sentiment analysis to determine the category of movies. Sentiment analysis, also known as opinion mining, is a core branch of natural language processing (NLP) and has made significant progress in multiple fields in recent years. Its aim is to identify and extract subjective information from text and determine its sentiment orientation [3]. According to the granularity of the text processed, sentiment analysis can be divided into word-level, phrase-level, sentence-level and document-level [4]. According to the category of the text processed, it can be divided into sentiment analysis based on news comments and sentiment analysis based on product comments [5]. The system it builds is used to identify and extract opinions in text, including effective identification of views on an article and classification of positive and negative sentiment tendencies, etc. [6]

Emotion classification is an important research direction in text mining in the academic field in recent years. It mainly studies the implicit emotional tendencies of language text information on the Internet. Usually, in addition to identifying viewpoints, emotion classification can also be used to extract descriptive features such as polarity, topic, and opinion holder, among other aspects [3, 7].

Compared with the traditional emotional tendencies obtained through actual research methods, current emotional analysis is generally divided into two approaches: one is to establish an emotional dictionary, and the other is to apply machine learning methods. Both of these analyze the emotional colors and emotional features contained in the text information through data mining technology,

which can not only provide real-time emotional tendencies or themes of the text, It also avoids the drawbacks of high costs such as human and financial resources in on-site research [8, 9]. In recent years, sentiment analysis methods have shown potential in topic prediction by combining high-frequency word extraction and dictionary matching. In film review applications, Chinese text sentiment analysis can efficiently identify opinion holders and topic conjunctive words [4], while the crawler tools developed in the Python language have solved the challenge of collecting large-scale review data [10, 11]. However, these methods usually neglect to directly infer the movie genre from audience comments, resulting in genre prediction models relying on external metadata rather than user-perceived data.

Overall, although sentiment analysis has formed a relatively mature research system in the field of sentiment mining in film reviews, there are still obvious gaps in the prediction research of domestic film theme types in China [11]. Most current academic literature focuses on the research of sentiment classification based on sentiment dictionaries or machine learning algorithms, aiming to quantify the overall emotional tendencies of audiences towards films, or concentrating on the correlation analysis between user behavior and product recommendations, thereby optimizing the recommendation algorithms of streaming media platforms. However, these studies generally rely on external metadata of films as the basis for analysis, and rarely directly mine genre features from the review texts generated by audiences - that is, the theme elements of the film actually perceived by the audience and the accompanying emotional feedback, which is precisely the key perspective for understanding the essence of film genres.

The research value of this article is precisely reflected in the exploration of this blank field. By innovatively combining high-frequency word matching with sentiment dictionary analysis, the research has broken away from the traditional limitation of relying on external labels and instead focused on the implicit type expressions in audience comments. This analytical approach, which starts from the perception of the audience, reveals the intrinsic connection between genre elements and emotional responses, providing more solid technical support for genre creation analysis, audience preference insights, and personalized recommendation services in the film industry.

2. Experimental Methods for Movie Classification

2.1. Experimental Preparation and Process

To ensure the reproducibility of the experimental part of this article, the configuration information table of the computer used in the experiment is limited to Table 1 below.

Table 1. Experimental Computer Configuration Information

Configuration Name	Configuration Information
Operating System	Windows10
CPU	Intel Core i5 11260H
GPU	NVIDIA GeForce RTX 3050
Programming Language	Python 3.12.2

At present, the analysis methods of sentiment dictionaries and machine learning algorithms are more commonly used in academic research on sentiment analysis. The analysis of text sentiment tendency using the analysis method of sentiment dictionary mainly involves organizing and classifying some commonly used sentiment words, assigning values according to different degrees of sentiment tendency to complete the construction of the sentiment dictionary, and then making judgments based on the scores of the words in the target text. Based on this principle, the experimental process of this paper is shown in Figure 1 as follows:

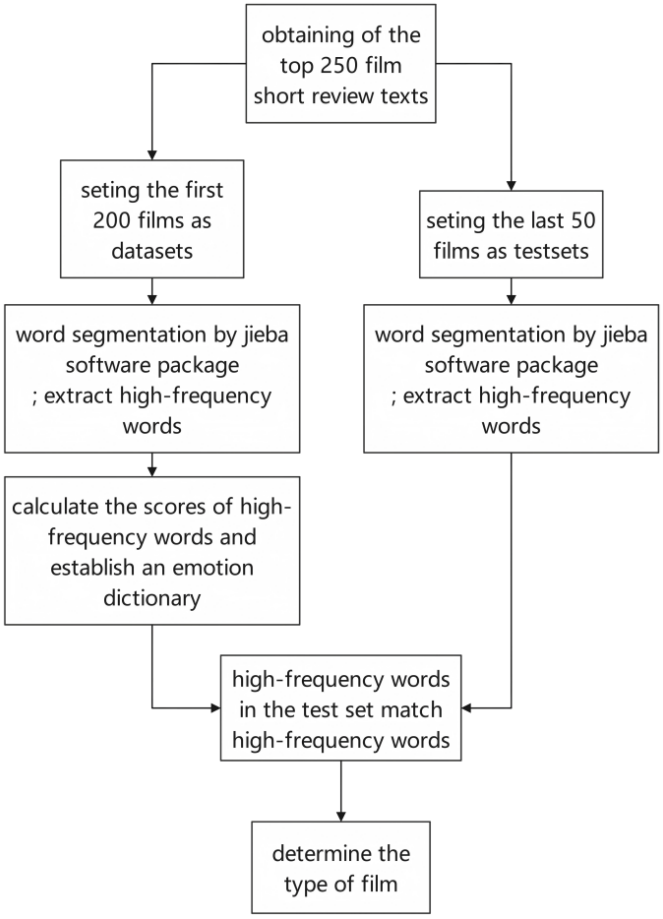

Fig. 1 Flowchart for determining film types (Picture credit: Original)

2.2. Data Acquisition and Processing

This article first uses the BeautifulSoup software package of Python to crawl the short reviews and classifications of the top250 movies on Douban. Among them, the first 200 films are used as the dataset, and the last 50 are used as the test set. For each film, about 5 pages of short reviews are crawled. The total number of short reviews in the dataset is approximately 20,000. The author considers this mainly because the number of short reviews for the vast majority of films is extremely large, mostly over 20,000, and it is not realistic to completely crawl them. Secondly, the ranking of short movie reviews on Douban is based on a certain algorithm used by Douban itself. The ranking of short comments is the result of a weighted average calculation of the votes of Douban members. By checking the ranking of short comments, it can be found that the number of useful short comments is basically decreasing exponentially.

After collecting sufficient short review texts, this paper uses the jieba software package to segment the short reviews crawled in the previous step. Because jieba has excellent word segmentation effects and advantages in Chinese word segmentation, this paper prioritizes the selection of the jieba software package, which is currently used more frequently and has a relatively better accuracy rate compared to other word segmentation technologies, for experiments. For the short comments after word segmentation, the author classifies the vocabulary, deletes the pause words, interjections, letter symbols, etc. that have little significance for judgment, and focuses on retaining the nouns, verbs, adjectives, etc. with high semantic content and keeps the frequency of their occurrence.

2.3. Experimental Methodology

After word segmentation, the author matched the high-frequency words of each film with the type of the film. For instance, in the film review of "Intouchables" on Douban, which is classified as a drama or comedy, the author identified "sympathy", which is considered a high-frequency word, as a

high-frequency word for both drama and comedy and saved it. In the case where repeated keywords in different movies need to be saved in the same type, the author combines the two words in the dictionary and adds their frequencies. From this, a dictionary of high-frequency words for 27 types of movies on Douban was obtained, and the effect is shown in Table 2.

Table 2. Movie Genre Dictionary (Partial)

Genre	High-Frequency Words (Partial)					
Crime	故事(362)	就是(245)	自己(221)	最后(208)	这部(160)	教父(153)
Drama	自己(1505)	故事(1177)	就是(1134)	我们(905)	最后(900)	不是(728)
Romance	自己(405)	爱情(347)	喜欢(307)	就是(266)	经典(247)	故事(240)
LGBTQ+	自己(74)	图灵(65)	李安(52)	爱情(46)	故事(44)	世界(41)
Disaster	丧尸(41)	爱情(37)	僵尸(30)	韩国(27)	最后(24)	人性(23)

Afterwards, the author crawled, segmented and extracted high-frequency words from the 50 films used as the test set in the same way as the dataset. In the subsequent stage of matching with the dictionary, the author calculated the score by comparing the frequency of high-frequency words in the test set movies with their occurrence frequency in a certain type in the dictionary, and then scored them. Each movie's type was determined based on its score in each type. The scoring formula is as follows:

For all the high-frequency words of a film $i=1,2,\cdots,m$ and a total of 27 film genres $j=1,2,\cdots,27$, there is the frequency of occurrence $h_i(i=1,2,\cdots,m)$ of the high-frequency word in the corresponding short reviews of the film, and each high-frequency word has its frequency of occurrence $d_{i,j}(i=1,2,\cdots;j=1,2,\cdots,27)$ in the total 27 film genre dictionaries. Thus, the probability P of a film in genre k is calculated by the following formula:

$$P_k=\sum_{n=1}^{m} h_m \times d_{m,k} \tag{1}$$

For instance, suppose in the movie "Rain Man", there are two high-frequency words: "road" (11) and "acting" (15) (the frequencies of the high-frequency words are in parentheses), and the frequencies of these two words in the "Drama" category in the dictionary are 39 and 52 respectively. Then, the score of "Rain Man" in the "Drama" category would be 11*39+15*52=1029 (points).

Finally, the author can calculate the type to which a film should be classified through its scores in different categories in one step. However, it is worth noting that since different movies on Douban belong to various categories, the algorithm used in the matching of this article is the top three types with the highest scores.

3. Results

After determining the types of test movies, this article will focus on the issue of the accuracy rate of discrimination. Let's first take a look at the distribution frequency of high-frequency words in different movie categories.

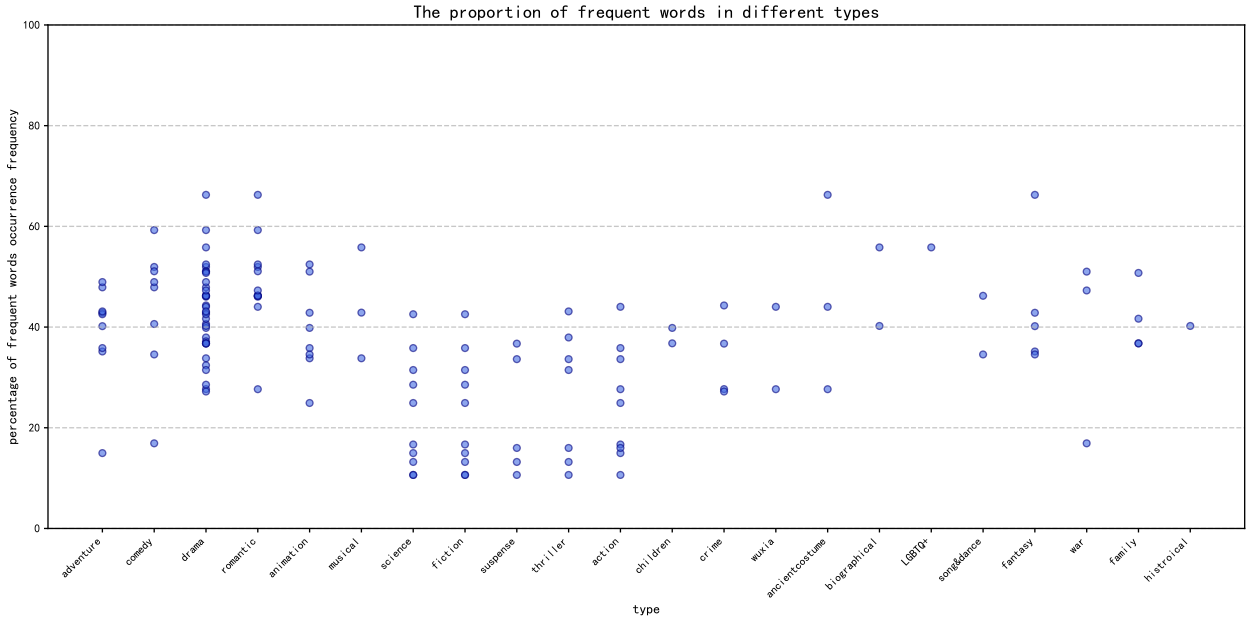

Fig. 2 The proportion of high-frequency words in different types (Picture credit: Original)

As shown in Figure 2, due to the varying popularity of different film themes in the market, there are significant differences in the distribution of high-frequency words among various types of films. For instance, high-frequency words account for a higher proportion in drama films, while those in historical and gay films are few and far between. Due to the limited content of the article, based on this distribution characteristic, this article selects some moderately popular film genres for analysis.

First, this article examines the experimental results of romantic films. Among the fifty films, the confusion matrix is (TP=12, FN=8, FP=0, TN=30). Based on this, the author calculated that the recall rate was 100%, and the model successfully identified all genuine and high-quality romantic movies. A precision rate of 60% indicates that when the model classifies it as a romantic film, it is only 60% reliable. An accuracy rate of 84% indicates that the overall prediction effect of the model is good, but it is affected by the imbalance of samples. An F1 score of approximately 0.75 indicates that the balance between the recall rate and the precision rate is above average. The above data indicate that the model is highly sensitive to romantic films but has a certain degree of accuracy imbalance. It also has a probability of misidentifying non-romantic films.

Next, this paper can focus on the experimental results of animated films. Based on the confusion matrix data of animated films (TP=5, FN=3, FP=3, TN=39), through simple calculation, it can be obtained that the correct determination rate of the model in animated films is 84%, while the recall rate and precision rate are both 62.5%. Compared with romantic films, although the accuracy rate has improved, false negative discrimination results have also emerged, indicating that further improvement is still needed in the discrimination of high-frequency words in animation.

Overall, the film theme prediction method proposed in this study which combines text sentiment analysis and high-frequency word matching can effectively identify film themes to a certain extent. This indicates that this method has certain feasibility and practicality, providing a new technical path for the analysis of film themes. Different from the traditional analysis approach that relies on external metadata (such as directors, cast groups, official type tags, etc.), it directly mines type features from audience comments, which is an innovative exploration of film type analysis methods. In terms of data sources, this paper uses the short reviews and film categories of the top250 ranking list on Douban. As one of the most influential film rating and review platforms in China, Douban has accumulated a vast amount of user-generated content. These data provide rich research materials for film theme analysis and emotion mining, ensuring the quality and representativeness of the experimental data. In terms of data processing, this paper uses the jieba software package to segment the crawled short comments. jieba has excellent segmentation effects and advantages in Chinese word segmentation and can accurately extract meaningful words. Moreover, for the short comments after

word segmentation, delete the pause words, interjections, letter symbols, etc. that have little significance for judgment, and focus on retaining the nouns, verbs, adjectives, etc. with high semantic content and keep the occurrence frequency of these words. This is helpful for more accurately extracting the high-frequency words related to the theme of the film. This analytical approach, which starts from the perception of the audience, reveals the intrinsic connection between genre elements and emotional responses, providing more solid technical support and unlimited possibilities for genre creation analysis, audience preference insight, and personalized recommendation services in the film industry, and facilitating the better development of the film industry.

4. Conclusion

Based on the data of short movie reviews on Douban, this study proposes a movie theme prediction method that combines text sentiment analysis and high-frequency word matching. The short review data of the Top250 movies on Douban was obtained through web crawling technology. The high-frequency words were extracted and an emotion dictionary was constructed for type matching. The experimental results show that this method can effectively identify the theme types of movies, especially demonstrating a high prediction accuracy rate in mainstream types.

However, the experiment also exposed several limitations, such as the relatively low accuracy and recall rate of the model in niche and other types reflecting the insufficient coverage of the high-frequency word bank for niche or specific types. The association mining between emotion and type features still needs to be deepened. Furthermore, the scale of the test set is limited, which leads to the problem that rare types cannot be effectively statistically analyzed due to the sparse samples of high-frequency words.

In future improvements, this paper still needs to increase the number of datasets and test sets to enhance the accuracy and universality of the model to strengthen the model's adaptability to diverse film themes. Secondly, the construction method of the sentiment dictionary can also be optimized to conduct secondary screening of high-frequency words of different categories more accurately, thereby enhancing the coverage and accuracy of type keywords. Finally, the author can also introduce deep learning models to capture semantic associations and contextual information in comments, or use NLP data augmentation methods based on large language models to enhance semantic understanding ability, and improve model performance.

Overall, this experiment verified the feasibility of directly mining genre features from audience comments, providing a new "user perception-driven" perspective for film theme analysis. In the future, through the iteration of technologies and methods, it is expected to further promote the precision and practicality of film genre research.

References

[1] Yu Yang. Sentiment analysis and topic extraction research on Douban movie reviews. Yunnan University of Finance and Economics, 2018.

[2] Ma Jia, Liu Yubang. Text mining and visualization data analysis of Douban short reviews on the program “Planting Season”: Research based on LAC sentiment analysis model. New Legend, 2025, (20): 73-75.

[3] Zhao Yanyan, Qin Bing, Liu Ting. Text sentiment analysis. Journal of Software, 2010, 21 (08): 1834-1848.

[4] Lu Wenxing, Wang Yanfei. A review of Chinese text sentiment analysis research. Application Research of Computers, 2012, 29 (06): 2014-2017.

[5] Liu Yulin, Jian Lirong. E-commerce online review data mining based on text sentiment analysis. Statistics and Information Forum, 2018, 33 (12): 119-124.

[6] Zhang Lu. Sentiment analysis of overseas readers’ reception and evaluation of Chinese translated literature: A case study of the English version of The Three-Body Problem based on Python sentiment analysis. Foreign Language Research, 2019, 36 (04): 80-86.

[7] Yan Chengxiao. Short text opinion extraction and polarity word analysis based on NLP. Software, 2022, 43 (03): 121-123.

[8] Qu Xiaoxiao. Research on clustering analysis and visualization methods of movie review data. Shandong University, 2018.

[9] Xu Delong, Lin Min, Wang Yurong, et al. A review of NLP data augmentation methods based on large language models. Journal of Frontiers of Computer Science and Technology, 2025, 19 (06): 1395-1413.

[10] Liu Zhiming, Liu Lu. Empirical study on Chinese microblog sentiment classification based on machine learning. Computer Engineering and Applications, 2012, 48 (01): 1-4.

[11] Zhou Zhonghua, Zhang Huiran, Xie Jiang. Sina Weibo data crawler based on Python. Journal of Computer Applications, 2014, 34 (11): 3131-3134.