ISSN: 2688-8653

Emotion Analysis of the Wandering Earth Film Critics based on SnowNLP

Xuan Wang

School of Shipping Economics and Management, Dalian Maritime University, Dalian, Liaoning, 116026, China

Abstract

Under the background of Internet, users' online comments will determine the future box office of a movie, so it is especially important for filmmakers to grasp the emotional trend of comments in time. To solve this problem, this paper takes The Wandering Earth as the research object to collect relevant comment data. Firstly, the comments are segmented, word frequency is counted. Then, the SnowNLP class library is used to analyze the emotion of the comments, and check the emotional tendency of the comments. In order to understand the core vocabulary that users pay attention to, and at the same time establish LDA topic model to master the topic distribution of positive and negative comments.At last, Word2Vec is used to transform the text into word vectors for K-means algorithm clustering, and comments are aggregated on topics.The application of this model can provide reference for relevant practitioners in film criticism and analysis.

Keywords

Text Mining; Sentiment Analysis; LDA Topic Model; K-Means Clustering.

1. Introduction

In recent years, China's film industry has been developing vigorously. Among the major film types, science fiction films have always been one of the most popular types in the cinema of Generation Z. Among the sci-fi fans of Generation Z, The Wandering Earth, directed by Frant Gwo in 2019, has the highest topic. With the development of the Internet, many sci-fi fans will rely on the Internet to express their views on movies,The endless stream of user reviews affects the reputation of films, and users will decide whether to buy tickets to watch them according to the existing reviews of films. Therefore, film reviews are particularly critical for film producers. Grasping the emotional trend of film reviews in time is conducive to formulating corresponding price strategies in the film market.

At present, the mainstream film review research techniques include text mining, sentiment analysis and so on. Text mining refers to the process of extracting unknown, understandable and finally available knowledge from a large amount of text data, and at the same time using this knowledge to better organize information for future reference, which plays a vital role in making decisions for enterprises.Emotional analysis, also known as opinion mining, tendentiousness analysis, etc. Based on text data, it is the main content of natural language processing (NLP). In short, it is the process of analyzing, processing, inducing and reasoning subjective texts with emotional color.

Ye and Wang et al.[1]took the American SciTS conference text as a case study, extracted the themes from the text, mined the emerging themes in the field and constructed the time series of emerging themes, which reflected the important significance of text mining in emerging fields. Ye and Wu et al.[2]used emotion analysis technology to identify medical service quality themes and their emotions from online patient reviews,The emotion recognition model of service quality theme based on LDA and BiLSTM model effectively improves the patients' satisfaction with hospitals and doctors. Cao and Li[3]to extract the online comments of consumers of electronic products on the e-commerce platform, and then classify the emotions to get the products that consumers like, and provide valuable information for the corresponding enterprises.

All the above studies show the important role of text mining and sentiment analysis in the era of big data, and more of them focus on the continuous improvement of the model, lacking the detailed analysis of the sentiment analysis results. At present, film and television practitioners have the need for emotional analysis of film reviews. Firstly, this paper preprocesses the collected data,Then, according to the data, word frequency statistics are carried out to get the core words that users pay attention to, emotion analysis is carried out by using SnowNLP, positive and negative comments are obtained, and then LDA topic model is used to analyze the positive and negative comments, and the subject words with the highest probability among the positive and negative comments are obtained. In order to show the aggregation of user comments more clearly, the features of the obtained text are extracted by Word2Vec,Then, the K-means algorithm is used for cluster analysis, and finally the aggregation of user comment topics can be clearly seen. The specific process is shown in Figure 1. In this paper, The Wandering Earth's film reviews are selected for text mining and emotion analysis, which will help film manufacturers to have an overall grasp of the film and make corresponding decisions in the face of the needs of users in the film market at any time.

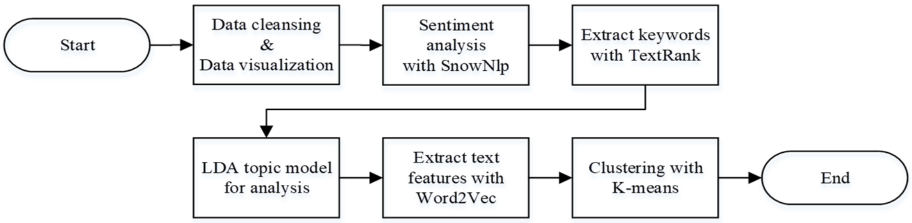

Start
Data cleansing & Data visualization
Sentiment analysis with SnowNlp
Extract keywords with TextRank
LDA topic model for analysis
Extract text features with Word2Vec
Clustering with K-means
End

Figure 1. The work flow chart of this paper

2. Data Preprocessing

2.1. Data Sources

The data of this article comes from the comment data generated by Douban from February to April, 2019, with a total of 480 pieces, including the location of users, user comments, user comments, the number of comments and other dimensional information.

2.2. Data Cleaning

Data type conversion, especially the format conversion of comment time, is performed on the obtained data, and at the same time, the useless data features of this article, such as user account level and other information, are removed, then the missing values are processed, the missing parts in the data are filled with -1 or unkonw, and finally the data reviewed is removed, that is, duplicate data is removed.Ensure the authenticity and validity of the processed comment set.

3. Establishment of Model

3.1. Emotional Analysis

3.1.1. Word Frequency Statistics

This paper selects the Chinese word segmentation component in python, which can perform Chinese word segmentation, part-of-speech tagging, keyword extraction and other functions, and supports custom dictionaries. Jieba word segmentation machine provides four word segmentation modes, including precise mode, full mode, search engine mode and Paddle mode, among which the more commonly used one is precise mode, which can cut sentences into the most accurate ones.It is suitable for text analysis. The specific use method is to call the cut () function, and the returned result is an iterator.

After word segmentation, regular expressions are used to remove non-text characters such as punctuation marks and numbers, and word frequency statistics are made for the text after the characters are removed. According to the word frequency statistics, word cloud maps are made to show the words with high frequency in comments. The higher the word frequency, the more important the users are, and the higher the attention they get.

3.1.2. Emotional Score

SnowNLP is a class library written in python, which can handle Chinese text content conveniently, and it comes with some trained dictionaries. Calling the sentiments property of the text can output the positive probability of the text. The closer the value is to 1, the greater the probability of the text being positive. The comment data obtained in this paper is called SnowNLP library to score emotion,The emotional score of each review is recorded, and the emotional scores of all the review data are statistically analyzed, which intuitively shows the emotional tendency of the film review.

3.2. Keyword Extraction

3.2.1. Extraction of Key Words

TextRank algorithm was originally used for automatic summarization of documents. It is an analysis based on sentence dimensions. Each sentence is scored by using TextRank, and then the N sentences with the highest score are selected as key sentences of documents to achieve the effect of automatic summarization. TextRank algorithm is a sort algorithm based on graph theory, which is often used in text processing. Its basic idea comes from PageRank algorithm of Google.The text is divided into many constituent units, including words, sentences, etc., and the importance of the text is sorted by voting mechanism. The key words can be extracted and summarized only by the information of a single document itself. This algorithm does not need to learn and train multiple documents in advance, so it has been widely used. In this paper, TextRank algorithm is used to extract TextRank keywords from the whole text , by outputting the Top20 keywords and their weights, you can know the general direction of user comments and grasp the dynamic information of users in time.

3.2.2. Analysis of LDA Theme Model

LDA(Latent Dirichlet Allocation) is a document topic generation model, also known as a three-layer Bayesian probability model, which includes three-layer structure of words, topics and documents[4]. The so-called generative model means that every word in an article is obtained through such a process that a certain topic is selected with a certain probability, and a certain word is selected from this topic with a certain probability.Finally, the document to topic obeys polynomial distribution, and the topic to word obeys polynomial distribution. It is a kind of unsupervised machine learning, which is often used to identify the hidden topic information in large-scale document sets or corpora. The bag-of-words method is adopted, which treats each document as a word frequency vector. Furthermore, the text information is transformed into digital information which is easy to model, but this method also has limitations, that is, it doesn't consider the order between words. Although it simplifies the complexity of the problem, it provides an opportunity to improve the model. Each document represents a probability distribution composed of some topics,And each topic represents a probability distribution composed of many words. The specific workflow of LDA theme model is shown in Figure 2.

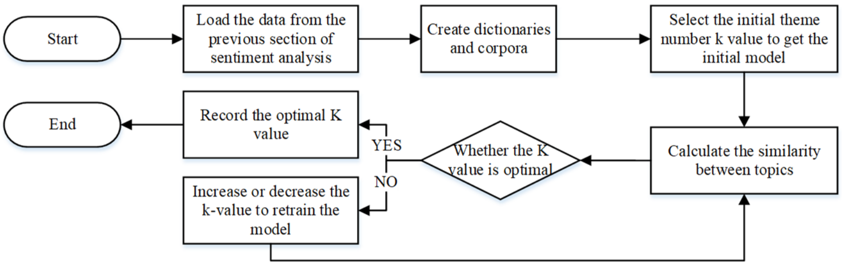

Start
Load the data from the previous section of sentiment analysis
Create dictionaries and corpora
Select the initial theme number k value to get the initial model
Calculate the similarity between topics
Whether the K value is optimal
YES
NO
Record the optimal K value
Increase or decrease the k-value to retrain the model
End

Figure 2. Workflow of LDA theme model

3.3. Text Clustering

3.3.1. Word2Vec Extracts Text Features

In this paper, Word2Vec method [5] is used to extract the features of text, which is a deep learning model based on prediction. Its basic working principle is to construct word vectors according to the context of current words, generally including CBOW and Skip-gram. Because of the space limitation, the two models are not introduced too much here. In this paper, CBOW model is used to construct word vectors. The specific principle is shown in Figure 3.

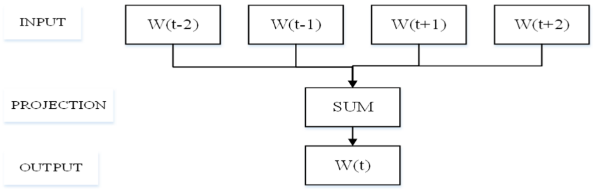


Figure 3. Working principle of CBOW model

W(t-2) and W(t-1) represent the upper information of the current word, and W(t+1) and W(t+2) represent the lower information of the current word, which are used as INPUTs for prediction in the input layer, and the word vector with the highest probability is calculated in the OUTPUT layer. In this way, 60 text features are extracted, which is convenient for the subsequent text clustering by K-means algorithm.

3.3.2. K-means Clustering

Clustering is to divide samples into several categories through the internal relationship between data without knowing any sample labels in advance. The core goal of K-means is to divide a given data set into K clusters, so that there is little difference between elements in the clusters, while there is a big difference between clusters. It shows that effective data classification can be divided into the following steps:

Step1 randomly selecting k objects from n data objects as initial clustering centers;

Step2 according to the average value of the objects in the cluster, reassign each object to the most similar cluster;

Step3 Update the average value of clusters, that is, calculate the average value of objects in each cluster;

Step4 Cycle Step2 to Step3 until the standard measure function converges.

4. Solution of the Model

4.1. Solution and Analysis of Emotion Analysis

4.1.1. Solution of Emotion Analysis

According to the results of word frequency statistics, make the corresponding word cloud map, as shown in Figure 4, which can more intuitively show the words with high frequency in the text. From the word frequency of the whole text, the high-frequency words include: China, science fiction, movies, the earth, and vagrancy. High-frequency emotional words include: sensational, human, hope,It can be seen that users' comments pay more attention to the film itself and see hope from the wandering earth. According to the score of SnowNLP after emotional score, the kernel density map of emotional score of data set is drawn, as shown in Figure 5. It can be seen from the figure that emotional scores moods are distributed at two ends, mainly around 0 and 1, with less distribution in the middle and the most around 1.It shows that the probability of positive comments is high, and most users' comments on the website are positive, so the film has a good reputation.

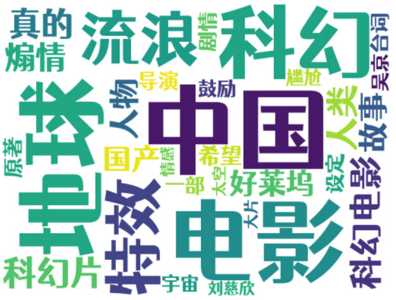


Figure 4. Word cloud picture

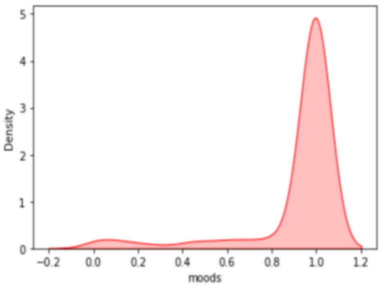
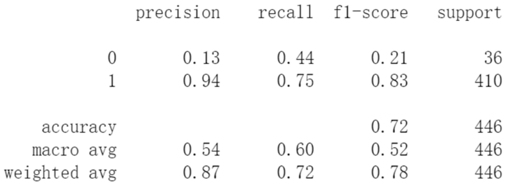


Figure 5. Kernel density diagram of emotional score of text

4.1.2. Analysis of Emotional Score

In this paper, the accuracy of SnowNLP analysis is evaluated, with 0.5 as the threshold, the emotion score is truncated by 0,1, and the confusion matrix is made. Confusion matrix can quickly analyze the misclassification of each category. This paper selects a 2*2 confusion matrix,The principle is shown in Table 1, in which TP(True Postive) indicates true positive, FP (False Positive) indicates false positive, FN(False Negative) indicates false negative, and TN(True Negative) indicates true negative. The final classification results are shown in Table 2, and the accuracy of SnowNLP sentiment analysis is analyzed according to the confusion matrix, as shown in Figure 6.



```
              precision    recall  f1-score   support

           0       0.13      0.44      0.21        36
           1       0.94      0.75      0.83       410

    accuracy                           0.72       446
   macro avg       0.54      0.60      0.52       446
weighted avg       0.87      0.72      0.78       446
```

Figure 6. Distribution of emotional scores

Table 1. Principle of Confusion Matrix

	Predicted value=1	Predicted value=0
True value=1	TP	FN
True value=0	FP	TN

Table 2. Confusion matrix results

	0	1
0	16	20
1	103	307

Figure 6 shows that the accuracy rate of emotion analysis based on SnowNLP is 72.42%, indicating that there are 410 positive emotion scores, while only 36 negative emotion scores, and the sample is extremely unbalanced. There are 307 right and 103 wrong judgments in category 1, 16 right and 20 wrong judgments in category 0. Overall, the accuracy of category 1 is higher.

4.2. Solution and Analysis of Keyword Extraction

4.2.1. Solution of Key Words

Output the Top20 keywords obtained by TextRank, and visualize their importance according to their weights. Finally, The Top20 are {'China', 'Movie', 'Earth', 'Sci-Fi', 'No', 'Wandering', 'Human', 'Sci-Fi Movie', 'People', 'Sci-Fi', 'Sensational', 'This', 'Hollywood', 'Domestic', 'We', 'Story', 'Hope', 'This', 'What', 'This'}.The word "sensational" occupies Top9, which means that the film is really sensational, and there are also "Hollywood" and "domestic". It may be that many comments have linked or compared this domestic film with Hollywood films."We" and "the story" should be that there are a lot of discussions on the plot in the comments, and the word "hope" also appears, which shows that many commentators have seen hope from the movies and presented a positive attitude.

4.2.2. Analysis of LDA Theme Model

After generating positive corpus and negative corpus, the number of topics is optimized, the model is trained by increasing or decreasing the number of topics, and the cosine similarity function is continuously calculated to find the optimal number of topics. Figures 7 and 8 respectively draw the average cosine similarity graphs of LDA topics optimized by positive and negative comments.

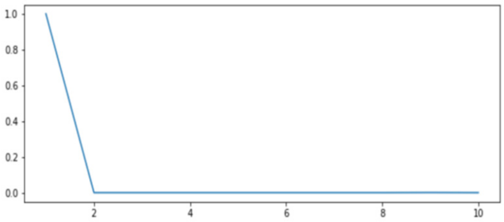


Figure 7. Positive comments LDA topics number optimization

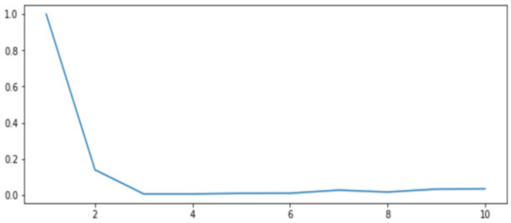
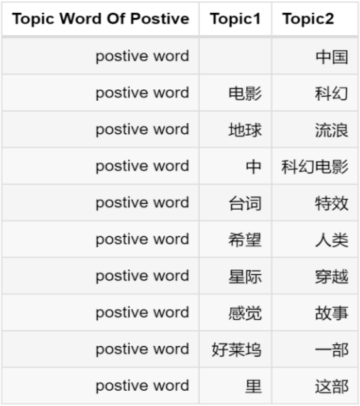


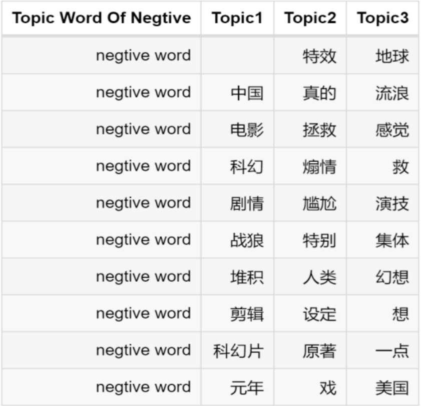
Figure 8. Negative comments LDA topics number optimization

It can be seen from Figure 7 that the number of topics suitable for positive corpus is 2, and from Figure 8 that the number of topics suitable for negative corpus is 3, so the subject words of positive and negative comments are visualized, as shown in figs. 9 and 10.



Topic Word Of Postive	Topic1	Topic2
postive word		中国
postive word	电影	科幻
postive word	地球	流浪
postive word	中	科幻电影
postive word	台词	特效
postive word	希望	人类
postive word	星际	穿越
postive word	感觉	故事
postive word	好莱坞	一部
postive word	里	这部

Figure 9. Key words of positive comments



Topic Word Of Negtive	Topic1	Topic2	Topic3
negtive word		特效	地球
negtive word	中国	真的	流浪
negtive word	电影	拯救	感觉
negtive word	科幻	煽情	救
negtive word	剧情	尴尬	演技
negtive word	战狼	特别	集体
negtive word	堆积	人类	幻想
negtive word	剪辑	设定	想
negtive word	科幻片	原著	一点
negtive word	元年	戏	美国

Figure 10. Key words of negative comments

As can be seen from Figures 10 and 11, there is less intersection between the three themes of positive comments and negative comments, indicating that the effect of theme separation is better. The first theme of the positive review is mainly the evaluation of the film, which is higher than the big frame of the human destiny. The second theme mainly includes the evaluation of the film's theme, feelings of rendering and its contents.As for negative comments, the first theme is similar to the evaluation of the big frame, but it is biased towards negative emotions; The second theme may be more biased towards the denial of the content of the film, and it also includes the criticism of the overall authenticity of the film, and the comparison between the two by associating it with American blockbusters, which leads to negative emotions;The third theme may be more inclined to dissatisfaction with film shooting technology, comments on editing special effects, and acting, which needs to be further improved and explored.

4.3. Solution and Analysis of Text Clustering

4.3.1. Solution of Text Clustering

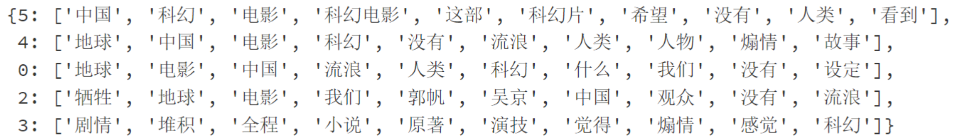
60 text features are extracted, and then the data is clustered by K-means according to the text features. According to the variance of the average score of the clustered effective categories, the K value is selected. It is verified that the variance of the effective categories grouped into 5 categories is the largest, indicating that there is a big difference between clusters at this time. The specific scores of each category are shown in Table 3.

Table 3. Average evaluation scores of various categories after clustering

	total_score	counts	ave_scores
5	2557.0	72.0	35.513889
4	7647.0	237.0	32.265823
0	2376.0	77.0	30.857143
2	1687.0	56.0	30.125000
3	389.0	16.0	24.312500

4.3.2. Topic Aggregation Analysis

The text of each topic is aggregated and keywords are extracted. The keywords in each cluster are shown in Figure 11.



```
{5: ['中国', '科幻', '电影', '科幻电影', '这部', '科幻片', '希望', '没有', '人类', '看到'],
 4: ['地球', '中国', '电影', '科幻', '没有', '流浪', '人类', '人物', '煽情', '故事'],
 0: ['地球', '电影', '中国', '流浪', '人类', '科幻', '什么', '我们', '没有', '设定'],
 2: ['牺牲', '地球', '电影', '我们', '郭帆', '吴京', '中国', '观众', '没有', '流浪'],
 3: ['剧情', '堆积', '全程', '小说', '原著', '演技', '觉得', '煽情', '感觉', '科幻']}
```

Figure 11. Keywords within cluster after clustering

Category 5: The average score is high, so it is positively connected. The key words are connected to see the hope of Chinese sci-fi movies, and the subject word-hope is mentioned. This topic not only mentions high-frequency vocabulary, but also talks about the hope and expectation seen from movies, which reflects the tolerance and encouragement to Chinese sci-fi movies.

Category 3: The average score is the lowest, only 24.31. This category is mainly about the movie plot, actors' acting skills, as well as the connection between the movie and the original novel. It may have found a huge difference between the two and made a spit.

5. Conclusion

This paper mainly analyzes The Wandering Earth from two aspects: emotional score of film reviews and topic extraction of reviews. In the emotional score, SnowNlp is used to score the emotion of the text, and TextRank is used to extract keywords in the comment topic extraction. Combined with K-means clustering method, five aspects that users pay more attention to are obtained, and the emotional trends of users are objectively and comprehensively grasped. Generally speaking,The overall user evaluation of the film shows a positive trend, which also confirms the high box office of the film. In the future work, K-means clustering algorithm will be further improved to enhance the degree of topic aggregation.

References

[1] Ye Guanghui, Wang Cancan, Li Songye. Identification and prediction of emerging topics in interdisciplinary scientific research collaboration based on SciTS conference texts [J]. Information Science: 1-10.

[2] Ye Yan, Wu Peng, Zhi Ming, Huang Wei, Zhang Liman. Research on online medical service quality identification based on LDA-BiLSTM model [J]. Information theory and practice: 1-10.

[3] Jihua Cao, Jie Li, Miao Yin, and Yunfeng Wang. 2022. Online reviews sentiment analysis and product feature improvement with deep learning. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (February 2022).

[4] Huang Yue, Zhang Xin. Research on Knowledge Structure Recognition Based on Subject Words and LDA Model [J]. Modern Information, 2022,42(03):48-56.

[5] Zhang Yuelin. Research on feature extraction and text classification of online goods based on Word2Vec [D]. Wenzhou University, 2019. doi: 10.27781/d.cnki.gwzdx.2019.200000000075.