2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) | 978-1-7281-9638-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/MLBDBI51377.2020.00016

Research on Chinese Movie Reviews Based on Latent Dirichlet Allocation Topic Model

Lan Zhao[1*]
[1]ZHEJIANG GONGSHANG UNIVERSITY
SCHOOL OF STATISTICS AND MATHEMATICS
Hangzhou, China
[*]Corresponding author: 1031501110@qq.com

Qian Zhao[2]
[2]ZHEJIANG GONGSHANG UNIVERSITY
SCHOOL OF STATISTICS AND MATHEMATICS
Hangzhou, China

Yating Wang[3]
[3]ZHEJIANG GONGSHANG UNIVERSITY
SCHOOL OF STATISTICS AND MATHEMATICS
Hangzhou, China

Abstract: **With the rapid development of the Internet, information acquisition has become more accessible, and a large number of online movie reviews cover the characteristics of audience preferences in the movie consumption market. Taking the top 10 popular movies in the Chinese box office as an example, this paper collects hot reviews, counts word frequency and visualizes the word cloud to show the keywords. At the same time, this paper uses the topic model to summarize key topics, so that we can gain insight into the hot spots. Finally, we conclude that the audience's attention mainly focus on the movie's theme, plot content, scene design, editing technology, character shaping, etc. Directors and actors have received a lot of attention, and the attention of domestic movies has increased significantly.**

Keywords: topic model; latent dirichlet allocation; short reviews

I. INTRODUCTION

In the process of the rapid development of the Internet, big data has played an important role in promoting human society. It is clear that the big data is complex, besides, it also has various kinds of types, fast processing speed and low value density. These characteristics equip the big data with the potential to become the equally precious resources as oil and gold mine in the next stage of social development. Nevertheless, how to extract the potential value from massive data has become the focus of people's attention. In addition to the most well-known structured data, text data in unstructured data also has great application value: learning customers' emotion through complaints, personalizing recommendation with the help of user comments [1], analyzing favorable scores of a brand or policy through Internet public sentiment monitoring [2], obtaining user's reviews through Web Crawler Technology, etc. In some way, text mining can provide us with different kinds of reference information. In addition, with the rapid development of the movie industry, Chinese audience's preference is the key factor for the gap between the domestic box office and the North American box office. According to the white paper of MOVIE.WEIBO.COM in 2019, the Chinese box office reached ¥64.266 billion in 2019, an increase of 5.4% over the preceding year. As a consequence, the gap has been further narrowed, head-effect of the movie market continues to strengthen and the movie ecology system continues to upgrade. All kinds of new domestic movies not only lead to the breakthrough of Chinese movie box and inject new vitality into the Chinese movie market, but also radiate the development of related industries and become the core hub of the industry. In view of the massive online movie reviews, provided that we employ text mining technology and the topic model to analyze them, we can understand audience's basic concerns. Then it will be beneficial for the Chinese movie market to get reformed and promote the further development of domestic movies.

II. SOURCE OF DATA

Douban is the most distinguished and prominent scoring website in Web2.0. In contrast, Maoyan is the most prevailing ticket purchasing platform. Additionally, it can provide immediate, precise and professional box office figures.

Considering that the movie scores on Maoyan are generally higher, the difference is insignificant, which provides little reference. While the scores on Douban are relatively objective and widely used, accordingly, we combined the two platforms to get data. By obtaining basic information of the top 10 movies in the Chinese box office from Maoyan, we use the Octoparse software to capture reviews of the top 10 movies from the Douban for further data processing and analysis.

III. METHODS

A. Word Cloud Visualization Technology

Keyword cloud mainly counts the high-frequency words in statistical documents, and takes the probability of the occurrence of the words as the basis for the final display of shape and size of the words.

The more frequently the words appear, the closer the words are to the center, the larger the font and the more eye-catching in the final lexicon. On the contrary, low-frequency words tend to be more marginal, smaller and less remarkable. Before the formation of word cloud, Chinese word segmentation is

978-1-7281-9638-1/20/$31.00 ©2020 IEEE
DOI 10.1109/MLBDBI51377.2020.00016

required. Rwordseg and jiebaR are the most commonly used word segmentation packages. The former requires the use of rJava to call word segmentation tools Ansj. Relatively speaking, jiebaR provides more comprehensive functions in the actual process. It provides a variety of segmentation engines, which includes MPSegment, HMMSegment, MixSegment, etc. Thus, MixSegment has a better effect. On the basis of Chinese word segmentation, it is available to use Wordcloud2 package in R to get a word cloud.

B. *Topic Model*

Latent Dirichlet Allocation (LDA) is the most prominent method in the field of Topic Model, the core premise of which is the existence of a set of implicit variables and a set of characteristics generated by them [3]. Each observation contains a mixture of characteristics from a subset of these implicit variables. Combining similar documents according to the key topics or the mixture of topics can be regarded as some form of clustering. With the help of topic decomposition, high-frequency words and the relative size of clusters, we can summarize information about a specific set of documents. In the topic model, there are categories, and the probability of each category is fixed. When we have a vector of length that contains the count of each category, we can estimate the probability of each category by dividing each item of the vector by the sum of all counts [4].

$$P(\text{word}|\text{document}) = \sum_{\text{topic}} P(\text{word}|\text{topic}) * P(\text{topic}|\text{document}) \quad (1)$$

Supposing that we want to predict the number of times each category will appear in a series of N trials. Provided that we have two categories, then we can use the binomial distribution to model this question. For K categories, the binomial distribution is generalized as multinomial distribution, in which the probability of each category is definite and the sum of probability of all categories is 1. The probability of random selection for a K-category multinomial distribution is as follows.

$$P(\vec{x}|\vec{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\sum_{k=1}^{K}\Gamma(\alpha_k)} \cdot \prod_{k=1}^{K} x_k^{\alpha_k - 1} \quad (2)$$

Among them, the x item is a vector with K components, where x_k represents a multinomial distribution. And the vector α is also a vector with K components, which contains K parameters α_k of Dirichlet distribution. The final calculation is the probability of choosing a specific multinomial distribution when a given combination of parameters is known.

IV. EMPIRICAL ANALYSIS

TABLE I. TOP 10 MOVIES IN THE BOX OFFICE

Rank	Movie	Average Attendance	Release Time
1	Wolf Warrior 2	38	2017-Summer
2	Ne Zha	23	2019-Summer
3	The Wandering Earth	29	2019-Spring Fest
4	Avengers: Endgame	23	2019-Weekend
5	Operation Red Sea	33	2018-Spring Fest
6	Detective Chinatown 2	38	2018-Spring Fest
7	Mermaid	44	2016-Spring Fest
8	My People, My Country	35	2019-National Day
9	Dying to Survive	26	2018-Summer
10	The Captain	26	2019-National Day

As we can see, “Mermaid” topped the box office’s top 10 movies with an average attendance of 44. By comparison, “Ne Zha” and “Avengers: Endgame” were at the bottom of the list. Through the correlation analysis, the Pearson chi square statistics of the box office and average attendance is -0.092, and the test result is not significant. It illustrates that the average attendance of a movie is little relevant to its box office. In terms of the release year and schedule, the movies in 2019 occupy half of the top ten in the list. Meanwhile, the competition among movies released in different schedules is also fierce.

A. *Descriptive Statistical Analysis of the Prospective Audience Profile Feature*

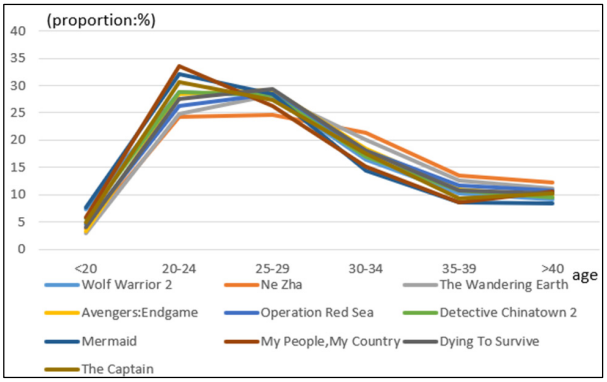

Figure 1. The age distribution of the prospective audience.

According to Maoyan data, Chinese movie consumption market covers a wide range of individuals of different ages. Among them, the post-90s audience is the main audience in the movie market, and the action movie is the main type of audience preference. On this basis, according to the top ten movies in the box office, the eye-catching comments of each movie are gained from Douban. And the main attention of audience can be further mastered by using text mining method.

B. *Word Frequency Statistics and Word Cloud*

First of all, the data preprocessing is carried out for the short comments crawled, including removing emoticons and simplifying the font. A user-defined thesaurus of stop words is constructed by combining the stop-words thesaurus from Harbin Institute of Technology, Baidu and Sichuan University’s Machine Intelligence Laboratory. With the help of Sogou thesaurus, the names and scores of the ten movies and relevant expressions, user’s dictionary is supplemented. By loading the jiebaR package in the R software, we segment the short comments. By loading the wordcloud2 package, the word segmentation text is displayed as a word cloud according to the

word frequency. Eventually, we take the top 100 words as an example to picture the word cloud according to their frequency of occurrence. Then we get a word cloud of the popular comments, which is as follows.

Figure 2. Word cloud of popular reviews for top 10 movies.

Among them, the key words of popular reviews include movie, story, director, China, character, Nezha, like, plot, sensationalism and role. From the point of view of comments, the audience's concerns can be summarized as plot content, directors and characters. As we can see, the audience's preference for domestic movies is also increasing.

C. *The Application of the LDA Topic Model*

First, according to the results of word segmentation, the word vectors are transformed into a list. Each line of the text is separated by spaces. Each line becomes a word vector stored in the list. Each word and word frequency are transformed into a cross table. Select the number of topics K= 5, 15, 20, then the results of the topic model are as follows [3].

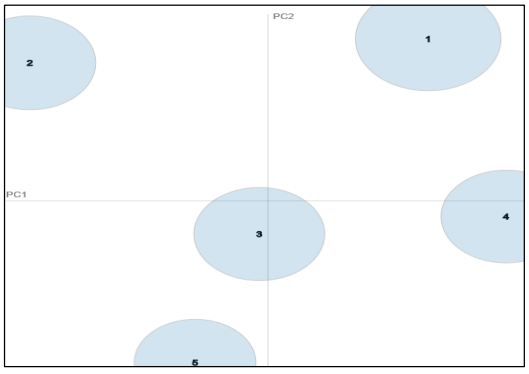

Figure 3. The results of the topic model at K=5.

Figure 4. The results of the topic model at K=10.

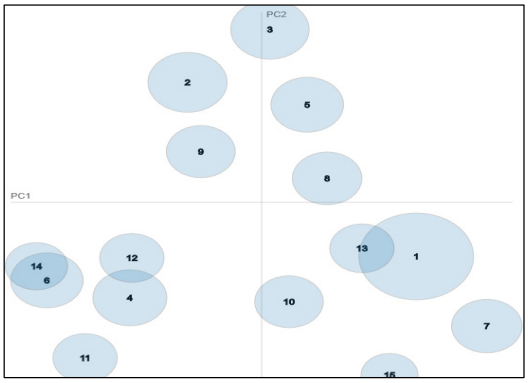

Figure 5. The results of the topic model at K=15.

The size of the circles in the figure indicates the frequency of the topic and the distance among the circles reflects the similarity among the themes, the formulas are as follows [4].

$$\text{saliency}(\text{termw}) = \text{frequency}(\text{w}) * \left[\text{sum\_tp}(\text{t}|\text{w}) * \log\left(\frac{\text{p}(\text{t}|\text{w})}{\text{p}(\text{t})} \right) \right] \quad (3)$$

$$\text{relevance}(\text{termw}|\text{topict}) = \lambda * \text{p}(\text{w}|\text{t}) + (1-\lambda) * \frac{\text{p}(\text{w}|\text{t})}{\text{p}(\text{w})} \quad (4)$$

In order to avoid the overlap of information between different topics, this paper selects K=5, where the lambda value can be adjusted to keep between 0-1. This paper sets lambda =0.3. Then get 5 topics and each topic contains 30 keywords. Take 25 words for example, the results are shown at the following table.

TABLE II. TOPICS AND KEYWORDS

Topic	Keywords
Topic1	Story, Scene, Emotion, Action, Picture, Shaping, Editing, Society, Time, Values, Ending, Process, Style, Chinese, Like, Fantastic, Worthwhile, Five Stars, Deliberate, Wolf Warrior, The Wandering Earth, Hello Beijing, Yuan Quan, Liu Haoran, Captain
Topic2	Domestic, Character, Science Fiction, Actor, Acting, Feeling, Performance, Blockbuster, Relation, Matters, Image, Original, Expective, Patriotism, Tension, Tribute, Home, Hope, Narrative, Reasoning, Business, China, Hollywood, The United States, Chen Kaige
Topic3	Plot, Figure, Technology, Spirit, Experience, Design,

Topic	Keywords
	Champion, The Eve, Meteor, Universe, Hero, Cinema, Patriotic, Real, Moving, Special, Surprise, Wonderful, Bad, Disappointed, Pity, Shock, Captain America, Iron Man, Zhou Xingchi
Topic4	Director, Audience, Screenwriter, Deng Chao, Wu Jing, Sacrifice, Regret, Funny, Meaningful, Encouragement, Series, Element, Circumstance, Way, Theme, Marvel, Motherland, Pace, Comedy, Performance, Work, Success, Box Office, One Star, Superhero
Topic5	Sensationalism, Lines, Lenses, Reality, The World, Type, Laughing Points, Wars, Ten Years, Humanbeing, Culture, Children, Strength, Hegemony, Spring Festival, Blood, Comparison, Future, Fans, Escort, Degree of completion, Xu Zheng, Zhang Yibai, Guan Xiaotong, Mekong River

The main features of Topic1 can be summarized as: stories, scenes, emotions, styles, editing; Topic2 can be summarized as: domestic, characters, science fiction movies, actors, matters, feelings; Topic3 can be summarized as: plots, characters, technology, spirits, design; Topic4 can be summarized as: directors, pace, screenwriters, topics, elements; Topic5 can be summarized as: lines, lenses, types, cultures, degree of completion. Topic1 focuses more on the stories and scenes, especially the styles of movie. Topic2 focuses more on domestic science fiction movies. Compared with Hollywood blockbusters, the production technology of domestic movies is also constantly improving. At the same time, the acting skill of actors are also very important, since the role shaping of actors can produce emotional resonance with the audience. In addition, combined with the list, patriotic movies are more attractive to the audience. Topic3 reflects the importance of plot design and character performance, which can promote the audience's emotion. Topic4 reflects the importance of the director and the theme. The director controls the pace and the scene design of the whole movie. The theme and elements of the movie are also essential to its success. Topic5 focuses on the memory points and empathy of the movie. Whether it is an impressive line or a shot, it can fully stimulate the audience's emotions and leave an aftertaste, and even the meaning is not enough. It is one of the highlights of the movie and can also improve the popularity of the movie.

V. Conclusion

Domestic movies are booming. The Chinese box office or the relevant products both show that domestic movies have already won the public's attention. Sometimes, movies released at different periods have certain characteristics and themes, since some themes are related to specified days, such as Spring

Festival, summer vacation and national day. With continuous improvement of the quality of script, acting skill of actors, scene design and production technology, the public are paying more attention to the domestic movies. Domestic movies have won a lot support, especially for patriotic movies, which are thought-provoking and inspiring. They not only publicize the local culture and depict history, but also awaken the public's sense of responsibility and mission of the times.

From the perspective of the audience, the focus is mainly on the type of theme, plot content, scene design, editing technology, etc. At the same time, directors and actors also attract much attention. The director needs to control the quality of the whole movie. The actor's acting skill is related to the role shaping and the transmission of the character's emotions. Besides, the fame of director and actor has a basic impact on the popularity of the movie. The overall pace and emotional communication of the movie are equally important. Whether the highlights and climaxes of the movie are brilliant enough, whether the central idea is clear, whether the movie is completed enough and meets the public's expectation will affect its reputation and popularity. Based on the opinions of the public, it is necessary for moviemakers to attach more importance to the theme, script quality, production technology, role shaping, etc. Then it will be conducive to firmly fit with the new trend of reform and development, which can improve the quality of domestic movies and create more fantastic works.

References

[1] Lin Jiang, Qilin Zhang. (2017) Research on Personalized Recommendation Strategy Based on Sentimental Analysis of the Reviews. Information Studies: Theory & Application, 40: 99-104.

[2] Renwu Wang, Jiayi Song, Chuanbao Chen. (2017) Application of Sentiment Analysis Based on Word2vec in Brand Awareness. Library and Information Service, 61: 6-12.

[3] Blei,D.M, Ng,A.Y, Jordan,M.I. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022.

[4] Kaveh,B., Hamed,N., Jeffrey,S. (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems With Applications, 127: 256-271.

[5] Lidong Wang, Baogang Wei, Jie Yuan. (2012) Document Clustering Based on Probabilistic Topic Model. Acta Electronica Sinica, 40: 2346-2350.

[6] Guangce Ruan, Lei Xia. (2019) Research of Topic Recognition Based on Term Weighting. Information Studies: Theory & Application, 42: 144-149.