Analysis of the Douban online review of the MCU: based on LDA topic model

Aiting Xu 1, Tianxing Qi 1* and Xinzhi Dong 1

1College of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang Province, 310018, China

*Corresponding author’s e-mail: 251568046@qq.com

Abstract. With the rapid development of the Internet and the movie industry, online reviews have shown explosive growth. Online review data contains a lot of valuable information, through text mining techniques such as the LDA topic model to analyze this information, to help the movie creators to understand the public's viewing needs, movie production team to reflect on the shortcomings in the production process, in order to promote the development of the movie industry to provide a reference. In this paper, the Douban online review data of the MCU are used as the research object, and the Chinese text is analyzed based on the LDA topic model. Firstly, data collection and preprocessing are carried out with various software. Then, the word cloud is used to visualize the core information in the online review data, and then the LDA topic model is used for a deeper semantic mining. The research results show that the rich and varied plot design, the audiovisual feast brought by advanced digital technology, the clever connection of bonus scene, the unique characters and actors and a super-hero's commercial movie mode are closely related to the success of the MCU.

1. Introduction

On the one hand, with the rapid development of the Internet and the rapid dissemination of information, more and more users tend to express their opinions, attitudes and emotions on the public platform. On the other hand, with the improvement of living standards, the movie has already changed from the extravagant enjoyment of the past to daily leisure. The number of online movie reviews has continued to increase in recent years. Hundreds of millions of data are scattered and are unruly text, but there is a huge hidden research value behind it.

Text mining is a cross-domain application that integrates information retrieval, information extraction, computational language, natural language processing, data mining and other technologies. Excavate hidden and valuable information in unstructured or semi-structured text data by using different technologies in multiple domains. The topic model is an important technology in the field of text mining. It can overcome the shortcomings of manual methods and allow automatic extraction of implicit semantic information from large-scale discrete data sets.

Douban is China's largest community of movie sharing and commentary. Not only the actual number of users who publish movie reviews is high, but also the speed of reading and publishing is relatively fast, and there is a certain degree of influence. And Douban movies can also provide a certain open interface, allowing large-scale automatic capture of movie reviews data.

The "Avengers: Endgame ", which was released on April 24, 2019, triggered a heated discussion of the public for nearly a month, and set off a new wave of watching. It can be seen that the MCU has great influence. As of April 9, 2019, the MCU has been released twenty-one movies, and the Douban
average score is 7.5. According to the statistics of China's box office network, the total box office in mainland China has exceeded 14.7 billion yuan. According to the statistics of Box Office Mojo website, the global box office has more than 18.5 billion US dollars, becoming the world's highest box office movie series. As one of the representatives of Hollywood commercials, the MCU is undoubtedly successful. The movie reviews reflect the public's attitudes and opinions on the movie. Deep excavation of hidden information in movie reviews is helpful to understand the reasons for its success in an all-round way.

In this context, this paper based on the LDA topic model studies the Douban movie review of the MCU. Crawling data through web crawler technology, and then the crawled text data are preprocessing, the analysis of word cloud and LDA topic model modeling, so as to deeply explore and analyze the movie review. Based on this, this paper understand the reasons behind the success of the MCU, which can help Chinese movie creators to accurately get the psychological needs of movie viewers, and provide a reference for the Chinese movie production team to compare and analyze problems. Then, some suggestions were made to improve the quality of Chinese movies.

The remainder of this paper is organized as follows. The relevant literature on the topic model is reviewed in Section 2. The research method is described in Section 3. Data collection and preprocessing are performed in Section 4. Section 5 describes the analysis of movie reviews. Finally, a number of suggestions were made in section 6.

2. Review of literature

The topic model is modeled according to the generation process of text data. By counting the co-occurrence information of words in different text data, the model can abstract the topic that is approximate textual meaning, so that the text data is represented as the topic form, which can greatly reduce feature Space Dimension of text data [1].

The earliest origin of the topic model was in 1990. The Latent Semantic Analysis (LSA) proposed by Deerwester et al [2]. LSA achieves the dimensionality reduction of words in the document by decomposing the singular values, and then maps the word to the hidden semantic range. The essence is to use the hidden semantic information between words. However, this method of analysis can neither explain the meaning of negative values due to the decomposition of singular values, nor the semantic information of polysemy, and topic is difficult to understand. In 1999, Hofmann [3] proposed the Probabilistic Latent Semantic Indexing (PLSI) by using probabilistic model to simulate the generation process of words in documents. Both documents and words are regarded as a random variable, and then potential topic variables are introduced. Although it can solve the problem of polysemy and multi-word synonym, it may produce over-fitting. Based on Hofmann's research, Blei et al [4] incorporated Bayesian ideas into the model and proposed the Latent Dirichlet Allocation (LDA) model in 2003. The model considers that the distribution of each topic is a multinomial distribution whose parametric compliance Dirichlet distribution. In 2004, Griffiths et al [5] further refined the LDA topic model by studying the distribution of $\beta$ parameters.

At present, the application scope of the LDA topic model is very extensive. On the one hand, it can deeply dig out the topic information hidden in the text, and summarize the valuable information in the text through the interpretation of the topic. On the other hand, because the mathematical logic of the model itself is very rigorous, various parameters are scientifically explained. The research around the topic model is mainly based on the LDA topic model. The focus of the research is to improve the existing problems of the model and to apply the model to solve practical problems.

3. Research methods

3.1. The basic idea of LDA topic model

Traditional methods of judging whether documents are similar are often limited to considering the proportion of the same word in different documents. However, this can easily lead to ignoring the semantic information implied in the word. Even if the same word between different documents is very
small, it is possible to have a high degree of similarity. This requires the introduction of the concept of a topic, which is a collection of a series of semantically related words that can effectively avoid problems in traditional methods. From a topical point of view, if the words in both documents appear in the same topic, it can think that the two documents are similar.

The LDA topic model is the process of document generation, which is essentially a three-layer Bayesian structure containing words, topics and documents. It can extract valuable information from the document set and efficiently process large-scale text data. The LDA topic model considers a document to be a collection of words, and there is no order between words. The process of generating each document is implemented according to certain rules: first select a topic from the topic distribution, then select a word from the word distribution in the topic, and then repeat the above two steps until the entire document is formed.

The core idea can be expressed by the equation (1) that the probability of a word appearing in a document is equal to the probability of occurrence of the word under a topic multiplied by the probability of occurrence of the topic in this document.

\[
P(\text{word} \mid \text{document}) = P(\text{word} \mid \text{topic})P(\text{topic} \mid \text{document})
\]  

(1)

The whole process is from words to topics to documents, and the theme is the information that needs to be mined. Since it is an unsupervised learning process, it is only necessary to know a document set and a certain number of topics to perform LDA topic model modeling.

3.2. The LDA topic model expression

A schematic diagram of the LDA topic model is shown in figure 1, where the squares represent the Chinese word segmentation results of the document, the circles represent the topics hidden in the document set, the arrows represent the relationship between the two variables, and the rectangular boxes indicate steps within this scope need to be repeated. The corresponding number of repetitions is shown in the lower right corner of the rectangle.

The specific steps of the LDA topic model are as follows: First, the topic \( t \) corresponding to each word is extracted from the multinomial distribution \( \theta \) corresponding to the document \( D \). Then, extract a word \( w \) from the multinomial distribution \( \phi \) corresponding to the topic \( t \). Repeat the above steps for a total of \( N_d \) times until each word in the document is extracted.

Figure 1. The LDA topic model diagram.

Suppose there is a document set \( DS \). Each document is marked as \( D \) and the entire document set has \( T \) topics. The word segmentation result of the document set is summarized, and the vocabulary is created, and the total number of words included is marked as \( V \).

The LDA topic model describes the relationship among documents, topics and words. Each document \( D \) in the document set \( DS \) corresponds to the multinomial distribution of the \( T \) topics, and is marked as the multinomial distribution \( \theta \). Each topic corresponds to the multinomial distribution of the \( V \) words in the vocabulary, and is marked as the multinomial distribution \( \phi \). The
multinomial distribution $\theta$ and $\varphi$ respectively exist a dirichlet prior distribution with hyper parameters $\alpha$ and $\beta$.

4. Data collection and preprocessing

4.1. Data collection

Douban movie reviews are in two forms, shorter reviews limited to 350 characters and longer reviews unlimited characters. The maximum number of characters that can be processed by a computer limited to a certain configuration is fixed. In order to collect more user reviews, shorter reviews are selected as the data required for this paper. This paper uses Python software to collect data. Until March 31, 2019, there were more than 2 million short reviews on 21 Marvel movies that have been shown. Due to the limitations of crawl time, data processing levels, and other objective conditions, it is not possible to study all the data. Moreover, the Douban movie will fold the short review which is not related to the movie or contains personal attacks, etc. The ranking of the short review is also the result of weighted average calculation according to the vote of the Douban members. Through the adjustment of the algorithm, the value of the short review can be better reflected. Therefore, this paper collects the short review of the top 500 for each movie to subsequent research, and finally gets 10,500 pieces of data. According to the word count, the total number of words is 550767, and the number of characters is 599,216.

4.2. Data cleaning and normalization

Although the crawling data has been initially screened by Douban, some of the movie reviews are pure punctuation, some meaningless emoticons or web links, and the text content is not available. There is also some textual content that conveys information that is not related to the movie itself. In order to process these invalid data, it is necessary to clean the collected data. Invalid data is cleaned by using related functions and manual methods in Excel. There are 9533 pieces of data after cleaning.

Due to different habits of language expression, so users often have strong arbitrariness in writing movie reviews, which leads to valid data after cleaning still have some problems. For example, some reviews use English, traditional Chinese characters or Internet language, and there are even mistyped characters. For the same character or actor, different users have different names. Therefore, the data is normalized using the simple function of Excel combined with manual modification.

4.3. Chinese word segmentation

Chinese word segmentation is the basis of computer processing in Chinese. The quality of word segmentation will directly affect the semantic mining of the next step. In the more commonly used open source Chinese word segmenters jieba and Ansj, the latter has better performance in terms of word segmentation accuracy, word segmentation speed, new word recognition, etc., so choose Ansj as a tool for Chinese word segmentation.

4.4. Removing stop words

The stop word is defined as the UHF vocabulary in the text that is not helpful for information retrieval and should be removed, such as "a, be, so". Its existence will greatly interfere with the entire text processing process. On the one hand, it contains too little valuable information. On the other hand, because it appears too high frequency, it will have a negative impact on mining other important vocabulary information, thus reducing the efficiency and accuracy of text processing. Currently, it is a common method to remove stop words by constructing a stop word list. By comprehensively comparing the three commonly used Chinese stop words list of Baidu, Harbin University of Science and Technology and Sichuan University Machine Intelligence Lab, it is found that the average effect of Baidu's stop word list on different corpora is the best. In the end, this paper chooses to use the Baidu's stop words list to remove the stop words. Using pre-processed data can improve the quality of the next step of data analysis.
5. Data analysis

5.1. Analysis based on the word cloud
The word cloud is a visual display of high frequency keywords in text data. By drawing a word cloud, we can have a basic understanding of the core ideas of text data. The pre-processed word segmentation data is subjected to frequency statistics and sorted according to the frequency from high to low. Regarding the frequency parameters used to generate the word cloud, the threshold is set to 100 by experience, that is, words having a frequency greater than 100 are used to draw a word cloud. The word cloud of the Douban online review of the MCU is shown in the figure 2.

Combined with the analysis of word frequency statistics and the word cloud, "good", "fine" and other keywords that express the viewer's emotional attitude tends to appear frequently, which shows that most of the audience rated the MCU well. The words "Iron Man" and "Captain America" etc have a higher frequency, which indicates that the characters created have received high attention from the audience. In addition, the keywords of actors such as "actor" and "Robert Downey Jr" also appear frequently, which indicates that except paying attention to the characters, actors also attract the audience. There are also many viewers who are interested in the plot, which is reflected in the high frequency of keywords such as "plot", "scene" and "story". Words such as "3D" and "special effects" reflecting visual effects and words such as "music" and "soundtrack" reflecting auditory effects are high frequency, which reflects the viewer's attention to the audiovisual effect. Unlike common movies, "bonus scene", "end", "series" etc express that the keywords associated with serial movies appear frequently, indicating that the coherence of the whole series of movies is also the focus of the audience. All in all, the characters, actors, plots, audiovisual effects, and coherence of the MCU have received attention from the audience.

Figure 2. The word cloud of the Douban online review of the MCU.

5.2. Analysis based on LDA topic model

5.2.1. Select the number of topics
Many studies have shown that the quality of LDA topic model modeling is closely related to the selection of the number of topics. Therefore, choosing an optimal number of topics is the focus of the LDA topic model application, and it is also difficult. The most commonly used method is to use Perplexity as the evaluation criterion and choose the model with the minimum Perplexity to determine the optimal number of topics. Although the minimum Perplexity can make the prediction ability of the model the best, it tends to choose a larger number of topics, which leads to higher similarity between the selected topics, resulting in problems that are difficult to identify the topic, and affecting the result of further analysis. After comprehensive consideration, select a certain range of values to make
multiple attempts, and determine the appropriate number of topics by artificial selection. The number of topics is marked as K. After referring to the relevant literature, K is taken as 15, 20, 25, 30 respectively, and LDAvis visualization results of the LDA topic model is shown in the figure 3.

Each circle in the figure 3 represents a topic, and the size of the circle represents the frequency of the topic. The distance between the circles represents the similarity between the topics. It can be seen that some topics are far apart, indicating that they are extremely independent; some topics are relatively close to each other, even overlapping, indicating that there is a highly similar between them. The appropriate number of topics, the distribution between the various topics should be reasonable. From the results of LDAvis visualization, it is necessary to minimize overlap between topics. According to the topic distribution of the topic models, the number of topics of the LDA topic model is finally set to 20.

![Figure 3. LDAvis visualization results (from top to bottom, from left to right: K=15, 20, 25, 30).](image)

5.2.2. Modeling result analysis

The LDAvis visualization results of the LDA topic model modeling are shown in the figure 4. Look over the word related to this topic by selecting different topics, ulteriorly summarize the meaning of
these words, and then summarize the meaning of the topic. Each topic shows 30 words separately. The relevance of each word to the topic is determined by the equation (2).

\[
\text{relevance(word } w | \text{topic } t) = \lambda \times p(w|t) + (1-\lambda) \times p(w) / p(t)
\]

(2)

This means that the correlation between the word \( w \) and the topic \( t \) is controlled by the parameter \( \lambda \). If \( \lambda \) is closer to 1, the higher frequency of the word, the more relevant they are to the topic. If \( \lambda \) is closer to 0, then distinctive word is more relevant to the topic. A suitable parameter value is chosen by continually adjusting, ie \( \lambda = 0.6 \). According to the probability order by the word generated in each topic, the top 10 word of the 20 topics are extracted as shown in table 1.

![LDAvis visualization result (K=20)](image)

**Figure 4. The LDAvis visualization result (K=20).**

According to the modeling results as shown in the table 1, some words that do not accurately reflect the corresponding topic are manually filtered out. Finally, the LDA topic model result of the Douban online review of the MCU is shown in table 2.

By analyzing the model results, this paper finds that the audience's focus on the MCU mainly focuses on the plot, audio-visual effects, special highlights, characters and actors, comprehensive evaluation and cognition. In terms of the plot, it includes scripts, funny elements, comic adapted movies, and memories of youth. A variety of different elements can attract different types of audiences, and it can also lead to heated discussions among different groups of people. Diversified content not only enriches the plot, but also invisibly enlarges the movie audiences. In terms of audiovisual effects, it includes visual effects, auditory effects, special scene setup, and viewing experience. The MCU fully utilizes advanced digital technology to enhance audiovisual effects. The cold war history, alien civilization and other scenes that cannot be photographed in reality and the bold ideas about the time and space can be truly presented on the big screen through data technology. In terms of special highlights, it includes creating suspense. Each movie in the MCU seems to be independent of each other, but in fact they are connected through the bonus scene. On the one hand, bonus scenes can create suspense, which will arouse the curiosity of the audience. On the other hand, it can also provide clues and more detailed descriptions to enrich the character image of the movie. In terms of characters and actors, female characters, male characters and actors are included. It not only displays that the
characters are distinctive, but also displays that the actors brilliantly shaped the characters which can resonate with the audience, which is conducive to cultivating the loyal audience, and easy to the unfolding of Marvel's world view. In terms of comprehensive evaluation and cognition, it includes commercial blockbusters, super heroism and overall evaluation. The good reviews such as classic commercial blockbusters and successful super hero movies coexist with the bad reviews such as the production process of the assembly line and boring, non-nutritious popcorn movies. But overall, most audiences still agree that the MCU is a wonderful superhero commercial blockbuster.

Table 1. Top 10 probability words of the 20 topics of the LDA topic model.

| Topic | Top 10 probability words |
|-------|--------------------------|
| Topic1 | Role, story, character, rhythm, emotion, director, audience, narrate, shape, logic |
| Topic2 | Character, villain, role, story, play, film, plot, shape, no, lead |
| Topic3 | Marvel, movie, superhero, like, series, popcorn movies, assembly line, commerce, success, type |
| Topic4 | Marvel, Ant-Man, joke, bonus scene, comedy, funny, interesting, world, relaxed, Marvel Universe |
| Topic5 | Action, scene, fight, play, ok, special effect, marvellous, good, design, plot |
| Topic6 | Captain Marvel, movie, Marvel, role, heroine, hero, superhero, female, chewie, grow up |
| Topic7 | Marvel's The Avengers, Marvel, movie, series, hero, expect, Captain America, superhero, hope, super |
| Topic8 | Technology, negro, politics, movie, Africa, culture, Black Panther, right, myth, actor |
| Topic9 | Special effect, Marvel, visual effect, visual, play, story, space, time, cool, inception |
| Topic10 | Special effect, play, ok, joke, awkward, rhythm, nice, popcorn movies, film, plot |
| Topic11 | Groot, like, music, Part I, joke, Star-Lord, hit song, acting cute, nostalgia, song |
| Topic12 | Bonus scene, finis, Marvel's The Avengers, trailer, surprise, joke, subtitle, expect, hot-blooded, excited |
| Topic13 | Movie, U.S.A, hero, comic, blockbuster, superhero, film, popcorn movies, heroism, adaptation |
| Topic14 | Villain, Thanos, lead, superhero, hero, excellent, universe, ending, powerful, enemy |
| Topic15 | Spider-Man, superhero, lovely, like, Iron Man, Ant-Man, good, villain, grow up, youth |
| Topic16 | Effects, ok, 3D, play, cinema, IMAX, good, fight, picture, special effect |
| Topic17 | Iron Man, Robert Downey Jr, like, film, villain, excellent, actor, china, charm, Scarlett Johansson |
| Topic18 | Captain America, Iron Man, Marvel's The Avengers, like, Thor, Hulk, black widow, Winter Soldier, good, Part I |
| Topic19 | Thor, Loki, Hulk, Edward Harrison Norton, like, Tom Hiddleston, Scarlett Johansson, Natalie Portman, figure, heroine |
| Topic20 | Movie, Marvel, play, boring, good, film, logic, screenwriter, series, anticipate |

Table 2. The LDA topic model result of the Douban online review of the MCU.

| Topic | Number | Topic content |
|-------|--------|---------------|
| Scripts | Topic1 & Topic2 & Topic20 | Story, narrate, logic, plot etc. |
| Commercial blockbusters | Topic3 | Commerce, success, assembly line etc. |
6. Suggestions  

Based on the above analysis results, this paper puts forward some feasible suggestions. First of all, the producers of Chinese movies should strictly examine the content of the script and create movies that are popular with the public. The script is the basis for movie production. The complete story logic chain, the three-dimensional vivid character image, the contradiction between the characters, and the transition of the plot smoothly and naturally, all need a good script to achieve. Therefore, the producers of Chinese movies must judge the merits and demerits of the scripts in order to create high-quality movies. Second, the movie production team should learn advanced digital technology and apply it to movie production. Digital technology allows moviemakers to freely handle the relationship between space and time, infinitely expanding the possibility of scene scheduling within the lens, and perfectly showing the screen effects of 3D and IMAX. The advancement of digital technology has turned the unique imagination of movie creators into reality, and the application of digital technology to movie production helps to improve the quality of movies. However, it is important to note that digital technology cannot be used blindly. It depends on the type of movie and the actual plot setting to decide whether to use digital technology and which digital technology to use. Third, the producers of Chinese movies should establish a holistic view of the movie, and the connection between the movies is reasonable. Although the pattern of the MCU is the same, Marvel as a series of movies is not limited to creating separate role for less unrelated plot. The MCU strives to create a complete world of ideas and a comprehensive view of the universe. This trait also makes each movie based on the same basic framework, and the series of movies is inseparable. The makers of Chinese movies should regard the various elements of the movie as a whole, and they can't be separated. Only by integrating the various elements can the audience feel that the content in the movie is real and credible. In addition, movie producers should pay attention to the match between actors and characters, and popularity is not the only criterion. At present, the role selection of Chinese movies is too concerned about whether actors are popular stars. Although popular stars can ensure the box office in a certain extent, there are still great risks in relying solely on fans. From the perspective of movie creation, this is also not conducive to the production of high quality Chinese movies. In the process of casting, the producer should comprehensively consider various factors such as acting, popularity, and schedule. Finally, Chinese movies must incorporate Chinese-style positive energy values and resonate with the audience. The superheroes in the MCU often have extraordinary abilities. In order to maintain justice, protect the people and save the world, they constantly struggle with evil forces. This is the individual heroism in the mainstream value orientation of the United States. As a result, most of the audience resonated with the core ideas of the movie. This resonance is no national boundaries, and positive values can be universally recognized. For example, the core of the thought of "walking with the earth" in "The Wandering Earth" broke through the American-style individual heroism, but a Chinese-style heroism with the feelings of home and country, and it was highly praised by foreign media. This
shows that the proper integration of Chinese movies into the Chinese mainstream values can also help to improve the quality of movies.

Acknowledgments
This paper was jointly funded by the first-class disciplines (category A) and the preponderant characteristic disciplines of Zhejiang Province (Statistics of Zhejiang Gongshang University).

References
[1] Xu, G., Wang, H.F. (2011) Development of the topic model in natural language processing. Journal of Computer., 34: 1423–1436.
[2] Deerwester, S., Dumais, S.T., Furnas, G., et al. (1990) Indexing by latent semantic analysis. Journal of the Association for Information Science and Technology., 41: 391–407.
[3] Hofmann, T. (1999) Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International SIGIR Conference New York: ACM Press., 50–57.
[4] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research., 3: 993–1022.
[5] Griffiths, T.L., Steyvers, M. (2004) Finding scientific topics. Proceedings of the National Academy of Sciences., 101: 5228–5235.