Analysis of Users' Emotional Attitudes to Vulgar Videos Based on Data Mining

Jinbin Yi[1], Yunting Miao[2]*, Yifan Tu[3]
1654597240@qq.com[1], 771022025@qq.com[2], 3525234622@qq.com[3]

College of Finance and Economics, Jiangxi University of Applied Science[1],

Innovation and Entrepreneurship College, Jiangxi University of Applied Science[2],

College of Finance and Economics, Jiangxi University of Applied Science[3]

Abstract. With the development of network video, Vulgar Videos with its unique video content and form quickly became popular, but in its early stage, there were some problems such as vulgar content and poor user experience. In order to optimize user experience, this study explores the emotional attitude of users when they participate in the Vulgar Videos based on data mining. By using Python language to write the data collection logic, run the program to crawl the microblog comment data about Vulgar Videos, use TF-IDF algorithm to extract keywords and make word frequency statistics, use SnowNLP library for sentiment analysis, build semantic network and LDA topic model, and carry out topic analysis and extraction. Meanwhile, Jieba database will be used for word co-occurrence analysis to further excavate and analyze users' emotional attitude. Through the above process, this study draws the following conclusions: Vulgar Videos not only have entertainment properties, but also social properties and information properties. From the perspective of emotion analysis, a large number of netizens hold a positive attitude towards Vulgar Videos, while a small number of users are dissatisfied with the video content.

Keywords: Vulgar Videos; Data mining; User's participation behavior; Sentiment analysis.

1 Introduction

Vulgar culture is an online subculture that has grown in popularity with short videos. Its representative-Vulgar Videos exhibit barbaric growth and rapid popularity, attracting the attention and engagement of users due to its extensive content coverage and rich and diverse formats. However, due to the lack of legal and moral constraints in the early development of Vulgar Videos, vulgarity, vice, and other content that deviates from mainstream values have been questioned and criticized by people from all walks of life. How to guide the positive development of Vulgar Videos and make them vibrant is a topic worth deep thinking.

2 Literature review

2.1 Vulgar Video concepts

Vulgar Video is a subculture in which vulgarity and short videos merge into each other. Scholars have studied it for a short time, and there is no consensus on their concept. Some scholars believe

that Vulgar Videos tend to be positive or neutral (Liu Yi, 2018)[1]. While some scholars believe that Vulgar Videos are a mockery of rural groups and rural lifestyle (Yi sha, 2019)[2]. Based on the aforementioned literature and materials, this study defines Vulgar Videos as follows: Vulgar Videos are short videos released on short video platforms and filmed in towns and villages, including shouty wheat, social rock and earthy melodrama. They have low production levels, lack innovation in content and have a strong rural flavor.

2.2 User's participation behavior

"User's participation" is a kind of interactive creation and interactive behavior experience based on the interactive mentality and the object of concern (Barki H &Hartwick J, 1994)[3]. Some scholars believe that user participation is the behavior of leaving messages, browsing and other interactive relations between users and other users during the process of participating in community activities (Joon Koh, 2007)[4]. In the process of participation, users autonomously use media to communicate, rather than passively accept the role of communication (CAI Wen, 2011)[5]. Thus, this study provides an in-depth analysis of the emotional attitudes of users in Vulgar Videos through user engagement behavior.

2.3 Sentiment analysis

Sentiment analysis is also known as opinion mining. In short, it is a process of analyzing, processing, concluding and reasoning the subjective text with emotional color (Zhao Yanyan, Qin Bing & Liu Ting, 2010)[6]. Some scholars have pointed out that the information on the Internet is increasing day by day, which contains a huge amount of information with emotion (Lu Wenxing & Wang Yanfei, 2014)[7]. Vulgar elements are abundant on Weibo, which can be used for data mining, and have a wide range of representational power, which is important for the study and analysis of user's emotional attitudes towards Vulgar Videos.

3 Mining and analysis based on Vulgar Videos micro-blog data

In this study, Python language is used to write a data collection program to crawl the comments data related to Vulgar Videos on Weibo, TF-IDF algorithm is used to extract keywords and perform word frequency statistics, SnowNLP library is used for sentiment analysis, semantic network construction, LDA theme model and Jieba library for theme analysis and word co-presence analysis. To realize the mining and analysis of commentary data information, this research data collected 6560 microblogs and 22,876 comments related to Vulgar Videos on Weibo platforms from 2016 to 2022. Through a preliminary screening algorithm, 2706 invalid microblogs were eliminated with an effective rate of 58.7 percent. Excluding 12,042 invalid comments, the effective rate was 52.6 percent.

3.1 Feature extraction of text data

The vector space model (VSM) (G.Salton,1975)[8]can transform the frequency and specific topics expressed by word elements in microblog comments and text into quantifiable structured data for function extraction. Because the m*n--dimensional document-term matrix (DTM) created by VSM scatters plenty of terms, this study will use the lexical item-inverse Document frequency (TF-IDF) algorithm to extract text attributes from DTM, as shown in the following

Table 1.

Table 1. Text feature model

File	like	Lovely	Vulgar	Video	…	Terms
File 1	W_{11}	W_{12}	W_{13}	W_{14}	…	W_{1m}
File 2	W_{21}	W_{22}	W_{23}	W_{24}	…	W_{2m}
…	…	…	…	…	…	…
File n	W_{n1}	W_{n2}	W_{n3}	W_{n4}	…	W_{nm}

3.2 Word co-occurrence analysis

In order to clarify the relationship between keywords, this study compiled a relevant algorithm to count the high-frequency words in the document, extracted the 17 words with the highest frequency to generate a 17*17 co-occurrence matrix, and used Ucinet software to carry out co-occurrence analysis on the generated co-occurrence matrix to generate the word co-occurrence network diagram. Fig. 1 shows the keyword co-occurrence network diagram:
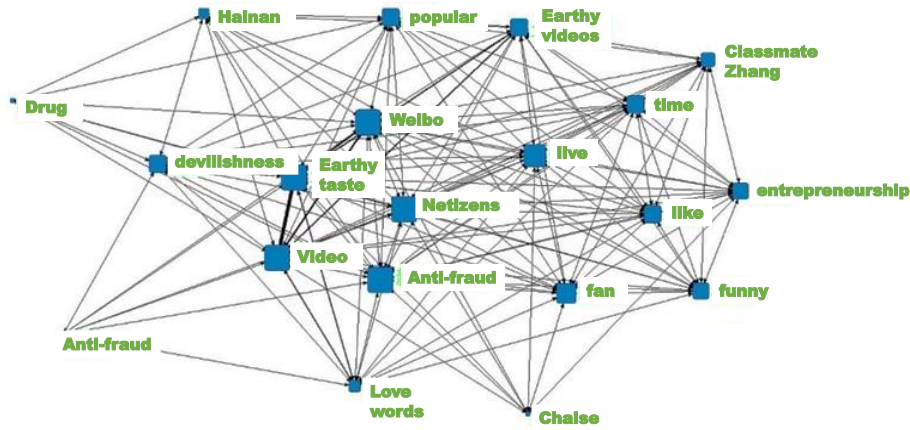

**Fig. 1**. 17 keywords appear in the network diagram

3.3 Topic analysis (LDA)

In this paper, we perform a topic analysis on the textual data of microblog comments. In this paper, the LDA model (Blei, 2003)[9] is used to approximate the parameter estimation using a special case of Markov chain Monte Carlo (MCMC) (Zhu Xinling, 2009)[10] algorithm Gibbs sampling. The prior functions α and β of Dirichlet function (Lin Yi & Li Jun, 2005)[11] are set as empirical values during processing:α=50/ K, β=0.1, and the algorithm is used to carry out iterative operation on the topic, and the line graph of confusion degree -- number of topics is obtained. It can be seen from the Fig. 2 that 10 is the best topic number, namely K=10.
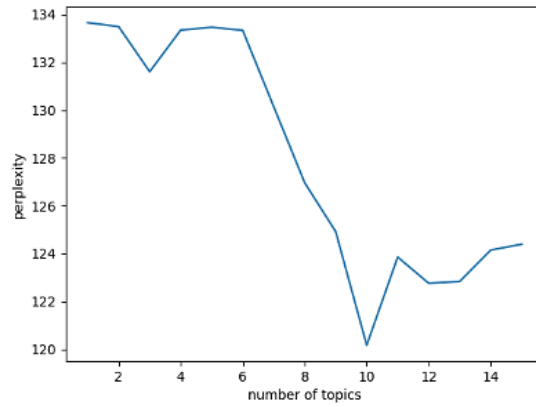
**Fig. 2**. Degree of confusion - number of topics line graph

3.4 Sentiment analysis

In this study, a Python program was written to directly call SnowNLP library for emotion analysis of microblog comment data, and according to the emotion score, a score below 0.4 was recorded as negative, a score over 0.6 was recorded as positive, and a score between 0.4 and 0.6 was recorded as neutral. Emotion analysis was conducted on 10833 comments after selection. Fig. 3 shows the emotion distribution histogram of the comments in Weibo.
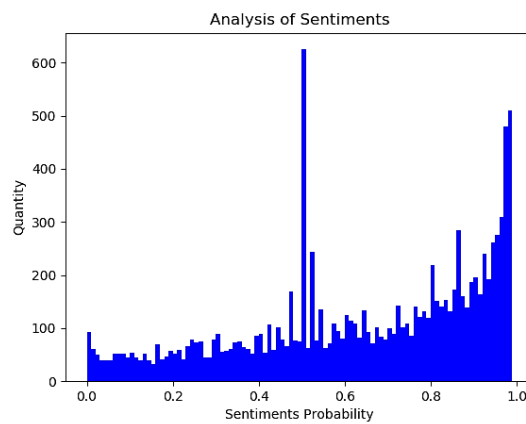
Fig. 3. Histogram of emotion distribution

Through the emotional and statistical analysis of 10,833 comments after screening, the statistical results can be obtained as shown in the Fig. 4 below. There are 6,297 positive comments, accounting for 58%, 2,218 neutral comments for 21%, and 2,318 negative comments for 21%. It can be seen that the vast majority of users have a positive attitude toward Vulgar Videos.
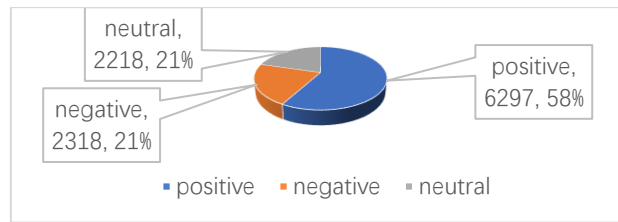
Fig. 4. Sentiment analysis statistical result graph

4 Conclusions

After crawling data related to Vulgar Videos in microblogs and comments and performing a series of text mining, it can be concluded that users are not only interested in entertainment, but also in life and news. It can be seen that Vulgar Videos not only have entertainment properties, but also occupy a highly significant position with information properties. In addition, the research also conducts an emotional analysis on Vulgar comments. Fifty-eight percent of citizens have a positive attitude toward Vulgar Videos, while only a few have a negative one, indicating that Vulgar Videos are more popular and preferred by people than Vulgar Videos.

Currently, the large amount of vulgar content in Vulgar Videos affects the emotional attitude of users while watching. Therefore, based on the conclusions of this study, the following recommendations are presented. First, creators should enhance the entertainment and information properties of Vulgar Videos and create videos that are more popular with people. Second, the platform should improve the accuracy of algorithmic push to meet the viewing experience and emotional orientation of different users. Finally, relevant administrative departments should strengthen supervision and introduce relevant policies to guide the development of Vulgar Videos.

This study, which analyzed the emotional attitudes of users in Vulgar Videos, still has shortcomings: First, the depth of the study is insufficient. Second, the sample data does not tell the whole story. The data mining objects in this study are single-valued, which does not completely reflect objective facts and does not guarantee the representability of the investigated data. To effectively address the above issues, the scope of data mining will be expanded in future studies to improve the credibility of the research.

References

[1] Liu Yi.: Vulgar is not local. Language construction. pp. 60-61 (2018)

[2] Yi Sha.: Fragmentation and Resistance: Rethinking the Phenomenon of China's Vulgar Culture--Based on the" Rural and Urban "Perspective. Audiovisual. pp. 240-241 (2019)

[3] Henri Barki, Jon Hartwick.: User Participation, Conflict, and Conflict Resolution: The Mediating Roles of Influence. Information Systems Research. pp. 422-438 (1994)

[4] Joon Koh,Young-Gul Kim,Brian Butler & Gee-Woo Bock.: Encouraging participation in virtual communities. Communications of the ACM(2). pp. 68-73 (2007)

[5] Cai Wen.: "From Facing the" Audience "to Facing the" Users "-- On the Impact of the Changes in Media Format on News Editing. International Press 33.05. pp. 993-1022 (2011)

[6] Zhao Yanyan, Qin Bing&Liu Ting.: Text Emotional Analysis. Journal of Software. pp. 1834-1848 (2010)

[7] Lu Wenxing, and Wang Yanfei.: Overview of Chinese Text Emotional Analysis Research. Computer Application Research. pp. 2014-2017 (2012)

[8] G. Salton,A. Wong,C. S. Yang.: A vector space model for automatic indexing. Communications of the ACM.pp. 613-620(1974)

[9] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. Journal of machine Learning research. pp. 993-1022 (2003)

[10] Zhu Xinling.: Review of Markov Chain Monte Carlo Method. Statistics and Decision. Statistics and Decision. pp. 151-153 (2009)

[11] Lin Yi, Li Jun.: Application Research of Dirichlet Function. Journal of Qingdao Vocational and Technical College. pp. 57-58,56. (2005)