International Journal of Business and Management Invention (IJBMI)
ISSN (Online): 2319-8028, ISSN (Print):2319-801X
www.ijbmi.org || Volume 11 Issue 1 Ser. II || January 2022 || PP 13-17

Research on Bilibili Corporate Account Video Comments Based on LDA Topic Model

Xin-yu Li
School of Management, Shanghai University, China

ABSTRACT: *In this study, we selected the video with the highest views of the Dingtalk official account on the Bilibili, obtained comment texts through crawlers, and established a comment text database including 57257 comments after data processing such as word segmentation. Use the LDA topic model based on the text database to analyze, conduct topic mining on the content of the video comment, and obtain the topic in the comment, so as to help the company to better carry out publicity activities.*
KEY WORD*: Video, Comment, LDA*

Date of Submission: 11-01-2022 Date of Acceptance: 26-01-2022

I. INTRODUCTION

With the development of the times, the publicity methods of enterprises have also undergone earth-shaking changes. Especially with the rise of the mobile Internet, the propaganda position of enterprises has also changed, from the original one-way transmission of television, radio, and newspapers, to self-media platforms such as Weibo and WeChat official accounts. The change of the propaganda platform also means the change of propaganda methods (Zaki and McColl-Kennedy, 2020). The company has transformed from a single information sender to an interactor who constantly communicates with consumers. Timely interaction with consumers helps companies better shape their own image and expand their influence (Yang et al., 2020). Among them, Durex is the leader in Weibo platform operations. With the popularization of mobile devices, lower tariffs, and faster Internet speeds, watching videos has become a way for most people to perform leisure and entertainment. The corporate propaganda position is also expanding toward video platforms such as Bilibili (B Station).

Bilibili was founded in 2009. It was originally an animation video website founded by anime enthusiasts, and has gradually grown into a larger UGC video website in China. Judging from the financial report for the first quarter of 2021, the average monthly active users of B station has exceeded 220 million, the average daily active users are 60.1 million, and the average daily usage time of users is as high as 82 minutes. From the perspective of age, the average age of users at B station is around 21, and the average age of new users is around 20. Therefore, how to use B station to promote the company's products and image has become a difficult problem faced by the company.

Many companies entered the B station early, but due to various reasons, few companies created popular videos, so it is difficult to achieve the purpose of publicity. Most of their video views are not high, and the number of comments is very few. Compared with professional video uploaders, the gap is huge. The video recommendation mechanism of B station is closely related to the interactivity of the video, and video comments play an important role in it. Therefore, it is particularly important to study the comments of the video.

During the COVID-19 pandemic last year, a video uploaded by Dingtalk on B station reached tens of millions of views. This video helped Dingtalk achieve promotion and rapid expansion in the office software market. Based on this video of Dingtalk, this research uses LDA topic modeling method to extract the topic of the comment content, so as to study the user comments.

II. LITERATURE REVIEW

2.1 Comments

Consumers' engagement to social media is usually measured by consumers' likes, shares, and comments. Peters et al. (2013) used reviews as part of their theoretical framework to assess the level of consumer brand promise. The research of De Vries et al. (2012) found that comments on brand posts reflect the popularity of the brand. In addition, posts that brands interact with can increase the number of comments. Some foreign scholars have conducted a series of studies on Facebook, Twitter, Youtube and their comments (Kwok and Yu, 2013; Alboqami et al., 2015).

Research on comments also tends to focus on the driving factors of comments (Li et al., 2019; Proserpio et al., 2020; Wei et al., 2021). The article by Rooderkerk and Pauwels (2016) examines the impact of

comment length, readability, and whether it is a question on users' comments on LinkedIn. Chandrasekaran et al. (2019) researched whether Facebook posts include links, photos, videos, and the information, entertainment, rewards, and social effects of the posts themselves on comments. Similarly, Barreto and Ramalho (2019) studied the influence of involvement and information on user comments.

Through literature review, it can be seen that most of the current research focuses on foreign-specific applications (Zhang, 2019; Chakraborty et al., 2021; Narang et al., 2021). While there are relatively few related applied researches on China, especially the research on B station comments. In these studies on reviews, they often only focus on the factors that affect reviews, including information, emotion, and entertainment, as well as whether they contain pictures, videos, and links, and most of them ignore the research on the reviews themselves (Liu, 2020; Zhong et al., 2020). This research hopes to supplement related fields through the study of reviews.

2.2 LDA Methodology Review

The topic model is a machine learning algorithm used to discover the main topics in the content of a large number of documents (Blei, 2012). Intuitively speaking, fitting topic models on text reviews can be seen as a reversal of the process of writing reviews. The topic model picks out word clusters that are often used together in comments, and determines the frequency of these words. Researchers can then use these data to come up with the most suitable topics for each type of word cluster. The parameters of the fitting probability distribution tell us the importance of each topic and the keywords related to each topic, which belongs to the probabilistic topic model. One of the specific methods is latent Dirichlet allocation (LDA).

There have been a series of studies using the LDA theme model. Some scholars have used the LDA model to collect low-dimensional themes of the phased policy texts of the free trade zone construction to help with decision-making research (Toubia et al., 2019); Online negative reviews conduct text mining to explore whether online businesses' requests for positive reviews will affect consumers' negative reviews (Yang et al., 2020); Zhang et al.(2019) conduct patent topic identification based on the industry chain perspective. The method of LDA topic model has been widely used in various fields of research.

In this paper, the LDA topic model method is applied to the review text mining of videos played by Dingtalk over 10 million. By extracting the theme of the comment text, we can study the audience's comment type, so as to help the company understand what kind of content will trigger the audience's stronger willingness to comment.

2.2 LDA Methodology Review

The topic model is a machine learning algorithm used to discover the main topics in the content of a large number of documents (Blei, 2012). Intuitively speaking, fitting topic models on text reviews can be seen as a reversal of the process of writing reviews. The topic model picks out word clusters that are often used together in comments, and determines the frequency of these words. Researchers can then use these data to come up with the most suitable topics for each type of word cluster. The parameters of the fitting probability distribution tell us the importance of each topic and the keywords related to each topic, which belongs to the probabilistic topic model. One of the specific methods is latent Dirichlet allocation (LDA).

There have been a series of studies using the LDA theme model. Some scholars have used the LDA model to collect low-dimensional themes of the phased policy texts of the free trade zone construction to help with decision-making research (Toubia et al., 2019); Online negative reviews conduct text mining to explore whether online businesses' requests for positive reviews will affect consumers' negative reviews (Yang et al., 2020); Zhang et al.(2019) conduct patent topic identification based on the industry chain perspective. The method of LDA topic model has been widely used in various fields of research.

In this paper, the LDA topic model method is applied to the review text mining of videos played by Dingtalk over 10 million. By extracting the theme of the comment text, we can study the audience's comment type, so as to help the company understand what kind of content will trigger the audience's stronger willingness to comment.

III. METHODOLOGY

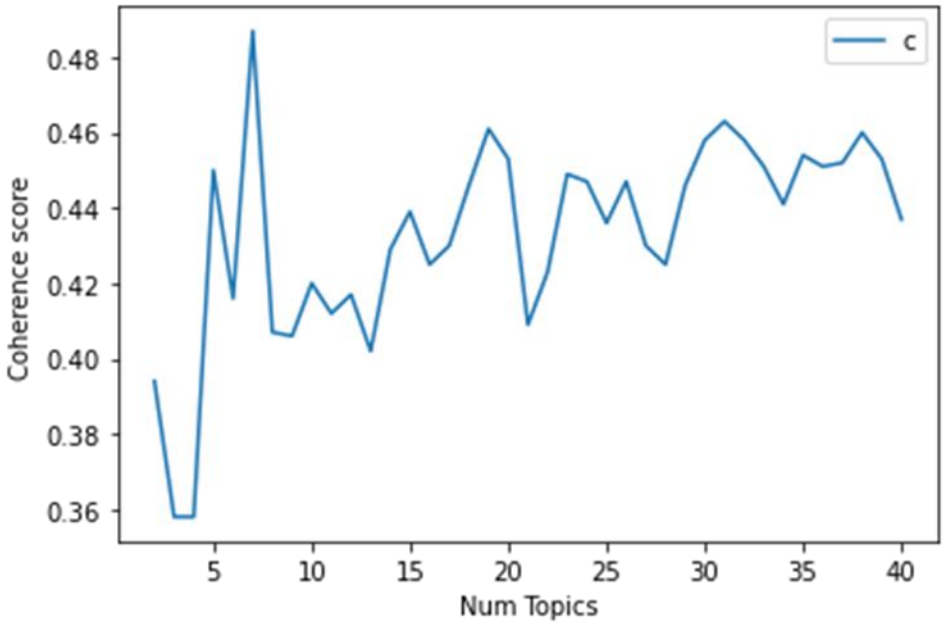
3.1 Data Source

The data in this study comes from the most viewed videos published by Dingtalk official accounts. The total number of views of the video reached 28.7 million, which is one of the most viewed videos among all official company accounts at B station. Research on the comments of this video has data collected from February 16, 2020 to April 21, 2021, a total of 60,659 comments. Programs are written in python, the content of relevant comments is obtained, the data is sorted and cleaned, and a total of 57,257 comments for analysis are finally obtained.

3.2 Comment Text Topic Mining Model

The research process is mainly carried out according to the following three steps. ①Collect relevant comments at B station to establish the comment text library needed for later analysis; ② Preprocess the text data, mainly including using the jieba library in python for text segmentation and removing useless stop words; ③Use the LDA topic model clusters related review texts, determines the number of topics, and judges different topic content according to the determined number of topics. The specific process is shown in Figure 1:

Figure 1: Concordance Score Distribution



IV. RESULTS

Input the processed text content into the LDA topic model. This study uses the LDA function in gensim in the python machine learning package for parameter training. Build a model by creating a dictionary and a bag of words. Set the reference document alpha to auto, the number of files considered at a time, chunksize to 100, and passes to 10.

The topic model can judge the quality of the topic model through the measurement of model fitting. This article changes the number of topics from 2 to 40, which allows us to check the model performance between the number of free topics from 2 to 40 (Mimno et al., 2011). The final result is shown in Figure 2. When the number of topics is 7, the Coherence Score is the highest, which is 0.487. Therefore, it is confirmed that the optimal number of topics is 7.

Since there are many feature words for each topic extracted through the LDA model, these excessive topic feature words are difficult to directly use for analysis, and the LDA topic model itself does not define the topic, and it needs feature words for induction. Therefore, this study selects the first 6 feature words related to each topic to summarize. The results are shown in Table 1 below:

Table 1: Topics and Feature Words

	Theme	Feature words
Topic 1	Positive emotion	Five star, Cute, Too, Pity, Praise, Happy
Topic 2	Software function	Dingtalk, Software, Teacher, Clock in, Class, Homework
Topic 3	Official begging	Five star, Dingtalk, High opinion, Begging, Study, Software
Topic 4	Ridicule	Official, Anti-addiction, Comments, Ding next time, System, Miserable
Topic 5	Video text	Nice, Dingtalk, Father, Child, Too, Miserable
Topic 6	Reply to begging	Definitely next time, Account, Advice, Dingtalk, Hour, Official
Topic 7	Negative emotion	Dingtalk, Next time, Taken down, No way, Get out, Store

The first topic is the positive emotions of the audience, which is mainly to evaluate the image of Dingtalk in the video and its humble attitude, including two emotions of cuteness and pity. In the video, Dingtalk asks students not to make negative comments again by begging and uses the anthropomorphic swallow image as the main body of the video, so it will produce cute and pitiful comments.

The second topic is related to software functions. During the epidemic, DingTalk was used by schools for online classes and DingTalk was used for learning. As a result, a series of experience perceptions related to student and teacher activities were derived. Therefore, this topic includes teachers, clocking in, class, and homework.

The third topic is related to the background of the release of this video. At that time, Dingtalk as a learning software was given negative reviews by students in major application stores, with one-star reviews, which caused Dingtalk's score to continue to decrease. Therefore, the official release the video asks student users to give a good review and give a five-star rating.

The fourth topic is the ridicule of this video. The anti-addiction and the next time Ding are all users' ridicules about Dingtalk. Anti-addiction is to hope that Dingtalk can limit the use time just like game software, and the next time Ding will be a certain homophony next time in Chinese. The main audience of anti-addiction is students, which is consistent with the cognition of the student group. In B Station, the "Geng Culture(Meme)" prevails, and the homophonic Geng is also widely loved.

The fifth topic is the evaluation of the video content. Father, child, miserable are all lyrics that appear in the video. The lyrics of the entire video are witty and humorous, using "father" to address the audience, using "child" and "miserable" to refer to the Dingtalk itself, which will be widely repeated in the comments.

The sixth topic is the response to the official rbegging for praise, including keywords such as next time, advice, etc. The video content sent a five-star praise request to the audience. Some audiences will seriously make advices based on the existing experience, while some audiences will respond to the request by saying the popular "next time" slogan in B station.

The seventh topic is purely negative emotions. Key words such as taken down, no way, and get out all reflect the relatively negative attitude of the audience towards the DingTalk software. During the epidemic, most students are on vacation and are psychologically dissatisfied with online teaching. Therefore, they will also have great dissatisfaction with related software.

V. RESULTS

5.1 Conclusion

Based on a user comment text data set of tens of millions of videos played on Dingtalk, this article analyzes the user comments of the video through text mining, uses the LDA model to identify the topic of user comments, and reveals the main topics of user comments. The main findings are as follows: video comments have 7 topics, namely positive emotions, software function, official begging, ridicule, video text, reply to begging, and negative emotions. Summarizing these seven aspects, comments are mainly divided into three types: emotional expression, event evaluation and mutual interaction. This requires companies to start from these three aspects when conducting publicity, to help them get more comments, so as to achieve the purpose of publicity.

5.2 Research contribution

This research has the following two theoretical contributions. First of all, this article analyzes the topic model through the comments of the video, which supplements the field of video communication research. Secondly, this article focuses on the official account's own creation and release of videos and supplements the research in the field of corporate publicity.

This study has the following management practice contribution. First of all, when dealing with negative events, companies should actively humorously resolve them, and use new communication tools to quickly carry out viral communication, to achieve the purpose of resolving crises and publicizing themselves; Second, when using new communication tools, such as a video platform, it must be more in line with the platform's tone. For example, at B station, a platform where young people are the main group, you must learn to use some creative methods in the platform or some stalks in the platform to create, doing so can often get very good results; Third, when companies create videos, they must be able to use event marketing to stimulate audience emotions and at the same time issue interactive requests to the audience. Only in this way can the audience's enthusiasm for comments be stimulated to the greatest extent.

5.3 Future research

This article also has certain limitations. First of all, because the video playback volume of the official company account is often not high enough, the research object is limited to a single video, and selection bias may occur. In the future, I hope to conduct research based on more high-volume videos; The second is the processing of the text. On platforms like B station, a large number of emoji and emojis will appear in the comment area to express the audience's attitudes and emotions, but there is no effective way to deal with these. In the future, I hope to conduct in-depth research on emoticons.

In the follow-up research, the latest deep learning technology will be used to improve the accuracy of text analysis through natural language processing, and further explore the potential value information in video reviews.

BIBLIOGRAPHY

[1]. Zaki M, McColl-Kennedy J R. Text mining analysis roadmap (TMAR) for service research[J]. Journal of Services Marketing, Emerald Publishing Limited, 2020, 34(1): 30–47.

[2]. Yang Y, See-To E W K, Papagiannidis S. You have not been archiving emails for no reason! Using big data analytics to cluster B2B interest in products and services and link clusters to financial performance[J]. Industrial Marketing Management, 2020, 86: 16–29.

[3]. Peters K, Chen Y, Kaplan A M, et al. Social Media Metrics — A Framework and Guidelines for Managing Social Media[J]. Journal of Interactive Marketing, 2013, 27(4): 281–298.

[4]. de Vries L, Gensler S, Leeflang P S H. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing[J]. Journal of Interactive Marketing, 2012, 26(2): 83–91.

[5]. Kwok L, Yu B. Spreading Social Media Messages on Facebook: An Analysis of Restaurant Business-to-Consumer Communications[J]. Cornell Hospitality Quarterly, 2013, 54(1): 84–94.

[6]. [Alboqami H, Al-Karaghouli W, Baeshen Y, et al. Electronic word of mouth in social media: the common characteristics of retweeted and favourited marketer-generated content posted on Twitter[J]. International Journal of Internet Marketing and Advertising, 2015, 9(4): 338–358.

[7]. Proserpio D, Hauser J R, Liu X, et al. Soul and machine (learning)[J]. Marketing Letters, 2020, 31(4): 393–404.

[8]. Li X, Shi M, Wang X (Shane). Video mining: Measuring visual information using automatic methods[J]. International Journal of Research in Marketing, 2019, 36(2): 216–231.

[9]. Wei Y "Max," Hong J, Tellis G J. Machine Learning for Creativity: Using Similarity Networks to Design Better Crowdfunding Projects[J]. Journal of Marketing, 2021: 00222429211005481.

[10]. Rooderkerk R P, Pauwels K H. No Comment?! The Drivers of Reactions to Online Posts in Professional Groups[J]. Journal of Interactive Marketing, 2016, 35: 1–15.

[11]. Chandrasekaran S, Annamalai B, De S K. Evaluating marketer generated content popularity on brand fan pages – A multilevel modelling approach[J]. Telematics and Informatics, 2019, 44: 101266.

[12]. Barreto A M, Ramalho D. The impact of involvement on engagement with brand posts[J]. Journal of Research in Interactive Marketing, Emerald Publishing Limited, 2019, 13(3): 277–301.

[13]. Zhang J. What's yours is mine: exploring customer voice on Airbnb using text-mining approaches[J]. Journal of Consumer Marketing, 2019, 36(5): 655–665.

[14]. Chakraborty I, Kim M, Sudhir K. EXPRESS: Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes[J]. Journal of Marketing Research, 2021: 00222437211052500.

[15]. Narang U, Yadav M S, Rindfleisch A. The "Idea Advantage": How Content Sharing Strategies Impact Engagement in Online Learning Platforms:[J]. Journal of Marketing Research, 2021.

[16]. Valluri C, Raju S, Patil V H. Customer determinants of used auto loan churn: comparing predictive performance using machine learning techniques[J]. Journal of Marketing Analytics, 2021.

[17]. Liu X. Target and position article - Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods[J]. Industrial Marketing Management, 2020, 86: 30–39.

[18]. Zhong N, Schweidel D A. Capturing Changes in Social Media Content: A Multiple Latent Changepoint Topic Model[J]. Marketing Science, 2020, 39(4): 827–846.

[19]. Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77–84.

[20]. Toubia O, Iyengar G, Bunnell R, et al. Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption[J]. Journal of Marketing Research, 2019, 56(1): 18–36.

[21]. Mimno D, Wallach H M, Talley E, et al. Optimizing semantic coherence in topic models[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, USA: Association for Computational Linguistics, 2011: 262–272.

Xin-yu Li. "Research on Bilibili Corporate Account Video Comments Based on LDA Topic Model." *International Journal of Business and Management Invention (IJBMI)*, vol. 11(01), 2022, pp. 13-17. Journal DOI- 10.35629/8028