Public Opinion Analysis About Online Teaching in the Era of Big Data: Mining and Analyzing Data from an Online Community in China

Xiaoqian Liu(✉)

School of Labor Relations and Human Resources, China University of Labor Relations, 45 Zengguang Road, Haidian District, Beijing, China
Liuxqian6251@sina.com

Abstract. Since the outbreak of the epidemic, a wide range of online teaching has had practical opportunities in China. Online teaching has also aroused extensive discussion among students, parents and public opinion. How to apply big data technology to understand the public opinion feedback on online teaching and improve the teaching effect has become an important issue in current education research. This research takes the network community of young people as the source of data acquisition. By initiating network requests, data extraction and cleaning, data storage, visualizing the stored data, and generating LDA topic model, this research has carried out public opinion analysis of online teaching, so as to provide support for the improvement of online teaching.

Keywords: Big Data · Online Teaching · LDA analysis

1 Introduction

After the normalization of epidemic prevention, online teaching has become regular needs. The education and teaching system is facing new challenges. The topic of online learning and teaching has become a hot topic of public concern. It is workable to obtain public opinion data through big data analysis. Based on this, we can objectively understand the social impact of online teaching. It can be said that public opinion analysis of online teaching is very necessary to promote the effect of online teaching.

Public opinion monitoring and analysis based on online teaching is a direct feedback on the effect of online teaching. In public opinion analysis area, there are different methods according to different needs. The common public opinion analysis methods for Internet data include network survey method, fuzzy recognition method and machine learning method. Various methods are applicable to different scenarios and achieve different purposes. The network survey method is carried out through the investigation and data analysis of the corresponding problems. The problems are more focused and targeted, but the scope and accuracy of the problem setting affect the analysis quality. Website traffic access analysis is a method based on statistical rules, which can understand the access of relevant resources through statistical technology; Some scholars

have also applied intelligent machine learning algorithm to the field of public opinion detection, mainly modifying the existing model based on the development process and characteristics of public opinion events to predict and predict public opinion [6]. Other scholars have deployed a system for monitoring employees' public opinion with Python [3]. These researches provide foundation for data mining in this study, but online teaching belongs to the new field of public opinion segmentation, and the mining and analysis of its specific public opinion situation is not sufficient.

At present, there has been a certain accumulation of text analysis methods in the public opinion system, which provides a technical basis for exploring the public opinion assessment of online teaching effect. In the discovery of public opinion hot spots, text clustering algorithm provides support for the discovery of public opinion hot spots in the weblog community. Sentiment classification method can be used for public opinion evaluation, K-nearest neighbor algorithm can be used for public opinion dissemination and tracking, and neural network and other algorithms can be used for public opinion early warning [4]. Other scholars have developed a topic emotion mining model with multi feature fusion based on the characteristics of Weibo [2]. The public opinion of online teaching has certain particularity, which is different from the general public opinion assessment. It is relatively scattered and has multiple topics. At present, the applied research of public opinion assessment and analysis system that can reflect the problems and influence of online teaching is still relatively lacking. Based on the characteristics of public opinion in online teaching, this study constructs a public opinion collection and analysis system for online teaching based on the data sources of typical online communities, and based on LDA topic analysis method to further find the focus of public opinion, so as to provide support for promoting the development of online teaching effect.

2　Research Design

2.1　Samples

In the period of China's extensive online teaching, all levels of teaching institutions actively explore and carry out online teaching. It is mainly the period of time after the outbreak of the epidemic in 2020. The most heated discussion and social feedback on online teaching discussions mainly come from communities where young people gather, of which "ZhiHu" community is the most typical. "ZhiHu" is famous for high-quality content and professional topics in many fields. It is gathered by young people. The information here appears in the form of questions and answers, and there are various discussions about online teaching. Taking "ZhiHu" community as the text source, this study captured the answers to all the questions related to "online teaching" in the early stage of the epidemic, and obtained 42739 answer data. By further structuring the answer data, Python is used to obtain various data information, such as user name, user profile, user gender, user ID, number of users' fans, answer content, answer time, answer comments, answer likes, etc.

2.2　Research Thinking

In terms of public opinion analysis, it often collects data in the corresponding online community through Python, and then analyses the crawled data through quantitative

analysis and sentiment feature extraction, so as to discuss the concerns of public opinion and the emotional projection of the public [5]. The first stage of this study is to apply this method to the data extraction and analysis of online teaching, focusing on solving a series of technical problems such as data source selection, data source structuring, data extraction and cleaning, data storage and data visualization. For the unstructured original data information obtained on the network, it needs to be structurally pre-processed, and rely on algorithm l and computer system automatic processing to further study.

After obtaining enough data, LDA analysis is used in the further analysis and focus of data. Among the existing studies, there is the research on online teaching network public opinion by searching "Weibo" keywords, and then the corresponding word frequency word cloud analysis and sentiment analysis are carried out [1]. However, the data of relevant situations in Weibo is much smaller than the sample size collected in this study, and the two network platforms are suitable for different social situations, Weibo is mainly self-statement, while "ZhiHu" is question response, and these situations are different. In order to make better use of the samples of this study and find the focus of public opinion for online teaching, this study adopts LDA topic analysis method, applies technical algorithms to find the core topic, and cluster the data.

The public opinion response of online teaching involves not only questions and answers, but also comments and opinions on answers. It is a public opinion system presented by cross interaction. Therefore, the continuous feedback of users on questions should also be taken into account when collecting data, which increases the difficulty of data collection. As the researchers pointed out, through the process of big data mining, intelligent information processing and semantic analysis, the network public opinion big data processing mode demonstrates and test the potential correlation, micro changes and dynamic evolution in public opinion analysis, and is used to assist decision-making [7]. On this basis, for the construction of online teaching public opinion analysis system under the network platform data, on the one hand, the characteristics and accuracy of the data itself are considered, on the other hand, the framework structure that can achieve the analysis goal is considered.

2.3 Implementation Process

In the specific implementation, Python language is used for data collection, structured data processing, data storage and data visualization. The specific steps can be described as follows: initiate network request ---- > data extraction and cleaning ----- > data storage ---- > visualize the stored data ---- > generate word frequency word cloud ----- > LDA topic model (Fig. 1).

2.3.1 Data Collection

This data collection mainly focused on ZhiHu web page and used Python language to capture data. Data capture can use the system technology of public opinion, but it is different from the method based on click through rate. Based on the analysis of the data structure of ZhiHu website, a complete set of crawling system is built for the website through python, which supports keyword input collection and date range input collection.
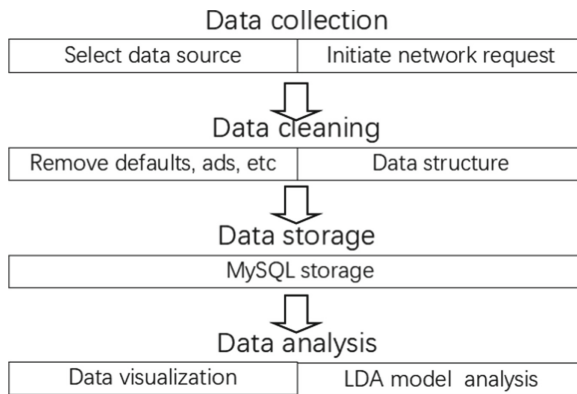
Data collection

Select data source	Initiate network request

Data cleaning

Remove defaults, ads, etc	Data structure

Data storage

MySQL storage

Data analysis

Data visualization	LDA model analysis

Fig. 1. Text data processing flow chart of "online teaching"

2.3.2 Data Processing

After the original data is collected, the data needs to be cleaned. The data collected in step 1 has many fields missing, or articles with obvious advertisements and other contents. Python's RE (Regular Expression) technology is used to check whether a string matches the required pattern, so as to filter advertisements and data with missing fields. It ensures that the remaining data is available, and carries out structured data processing.

2.3.3 Data Storage

Considering the need of real-time collection and the large amount of data, the structured data processed in step 2 is stored in the form of database, and MySQL database is used here. MySQL is a relational database management system, which saves data in different tables instead of putting all data together, which improves flexibility and later analysis efficiency.

2.3.4 Data Visualization and Analysis

The data visualization part mainly does the following work. The first is the word segmentation and word frequency statistics of ZhiHu's answers stored in the database. Here we use Python's JiebaDict and then for the answer content of ZhiHu, LDA topic analysis is carried out in order to find the potential words in the sampling document and find the relationship between two samples without the same words. It specifically shows what topics all the discussions focus on and what keywords these topics are related to.

3 Results

Under the impact of the epidemic, online teaching has become the main technical means to realize the non- suspension of classes. After collecting and analysing big data about online teaching in "ZhiHu" community, it is found that students are generally optimistic about online teaching, and teachers and students' maladjustment in teaching has become the highest focus of answers and comments, such as forgetting to turn off the microphone.
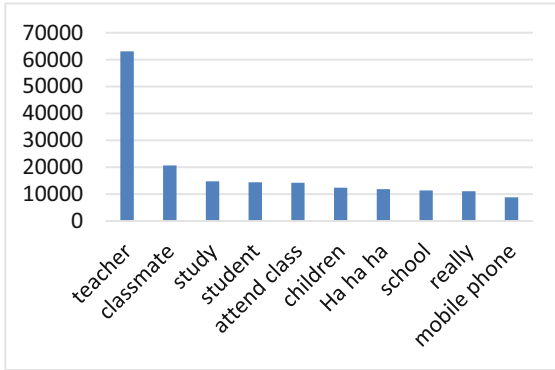
Fig. 2. Word frequency table about online teaching

3.1   Word Frequency and Word Cloud Analysis Results

In the word frequency analysis, words such as teacher, classmate, learning, student, class and child are in the top six of the overall word frequency ranking, which shows that the objectification characteristics in public opinion are obvious. Especially the frequency of "teacher" is 63116, while the word frequency of "student" is 14423. It can be seen that the current online teaching is still a teacher led mode, and the presence sense of students in online teaching is not high. On the other hand, it also shows that teachers are still in a dominant position in teaching. The new student-centered "online and offline" teaching mode emphasizes more on student-centered teaching. However, the current online teaching is still dominated by teachers' teaching, and students' participation in the classroom has not been enhanced. In the word frequency analysis, modal particles expressing happiness and positive attitude also appeared in the top 10 words, which can be seen that in general the attitude towards online teaching in public opinion is optimistic (Fig. 2).

In the data of questions and answers, the top three questions are "If you forget to turn off the microphone in online class, what should you do?" "What interesting things have you encountered in online class in winter vacation?" "Does the article (this screen may change fate) truly reflect the current situation of education? Is online class the way for education popularization?" It is obtained according to the descending order of attention to 570 problems. The problem of forgetting to turn off the microphone ranks first in terms of attention, pageviews and comments. The top seven responses with more than 35000 responses are embarrassing descriptions of forgetting to turn off the microphone. It can be said that the problems caused by forgetting to turn off the microphone have become the most concerned public opinion topic in the online community of young people.

In addition, according to the descending order of the number of view volumes, this question ranked secondly, "how to treat the case that girls in junior high school in poor households can't afford to buy smart phone for online classes, and then take a lot of drugs to send to hospital for rescue?" Which also triggered a lot of discussion about the impact of online teaching on social vulnerable groups.

3.2   LDA Analysis Results

LDA topic analysis model is a full probability generation model with clear internal structure, and can be calculated by using efficient probability inference algorithm. This study is based on the advantages of LDA model, through the tags of each answer sample, so as to find the question of text sentiment topic.

In the results based on LDA topics, a total of 20 groups of sentiment topic related phrases were obtained, and several typical topics were selected for introduction.

Firstly, we found a series of keywords for online physical education problems. Combined with high-frequency words, we can see that there is widespread confusion and discussion about how to improve online physical education and how to strengthen exercise through online classes. It should become a problem that educational institutions must seriously consider.

The probability presentation of keywords in topic analysis 6 also objectively presents online classes on homework and eye care. Combined with the text materials, we can see that the amount of homework, homework arrangement, learning time and the relationship with students' eye problems in online classes are also very noteworthy.

The clustering in topic analysis 10 shows the multiple participants in online teaching, which is different from the traditional teaching mode of teachers to students. Because the main place of online teaching is at home, parents are deeply involved in school education and classroom education. At the same time, different institutions and platforms are also trying to intervene in Online teaching in various forms.

The analysis in topic 12 shows to some extent people's discussion, suggestions and hopes for this new method of online teaching. If the suitable mode of online teaching is determined, how to improve it, and how to improve the knowledge absorption rate, it should become the focus of future education.

Table 1. Partial excerpts of LDA topic analysis results

Topic 1	0.073* "physical education class" + 0.035* "physical education" + 0.029* "school" + 0.025* "ranking" + 0.020* "Sports" + 0.018* "physical education teacher" + 0.013* "distress" + 0.012* "bad" + 0.012* "girlfriend" + 0.011* "attendance"
Topic 6	0.093* "homework" + 0.026* "eyes" + 0.020* "school" + 0.016* "exam" + 0.015* "teacher" + 0.012* "brush questions" + 0.012* "live broadcast" + 0.011* "math" + 0.011* "layout" + 0.011* "hours"
Topic 10	0.126* "children" + 0.032* "parents" + 0.031* "education" + 0.027* "institutions" + 0.024* "learning" + 0.023* "students" + 0.020* "Courses" + 0.017* "parents" + 0.017* "platform" + 0.016* "teaching"
Topic 12	0.098* "learning" + 0.021* "method" + 0.018* "suitable" + 0.016* "education" + 0.014* "suggestions" + 0.013* "ways" + 0.012* "Hope" + 0.012* "one" + 0.012* "effort" + 0.011* "knowledge"
Topic 14	0.021* "free" + 0.018* "teacher" + 0.016* "open class" + 0.014* "course" + 0.014* "video" + 0.012* "YouDao" + 0.008* "user" + 0.007* "teaching assistant" + 0.007* "audition" + 0.006* "learn"

The topic clustering obtained in topic 14 mainly focuses on the discussion of a wider range of free online education resources. In particular, many platforms are promoting online education through free audition classes and open classes. How to treat this emerging form of online education that complements school education has also become a hot topic of public opinion (Table 1).

4   Conclusion and Discussion

The requirements of epidemic prevention have put forward online teaching requirements for all kinds of schools. The current public opinion of online teaching is of great significance for the improvement and development of online education. For the analysis of online teaching public opinion, this study focuses on the question-and-answer online community "ZhiHu" gathered by young people, collected big data for data sources, and continued to analyse public opinion by means of word frequency analysis and sentiment topic analysis. Through big data technology, we tracked and summarized some topics that are crucial to the development of online education, including the problem of personal life information infiltration in teaching, the inability of poor families to buy equipment for online classes, and the trends of these problems' public opinion. And through the analysis of LDA model, we found more meaningful topics to be concerned, including how to solve the problem of physical education classroom organization when only online teaching is available; How to give consideration to children's personal life and eye health in online teaching; the deep participation of parents brought by online classes and the demand for the complex interaction mechanism of multiple educational subjects; the expectation of the public for online education; the online teaching development mode for other educational institutions, etc. These problems should be considered by current educational researchers.

However, in terms of data acquisition and analysis, there is still room for further expansion. In the analysis of online education public opinion, we can also collect big data information more widely, involve more social media interaction platforms, and apply more methods to the sentiment topic mining of short texts. In the field of public opinion tracking in the development of online teaching, there is still a broader research space, whether in the real-time assessment of new data or in more in-depth technology application.

References

1. Guo RR, Min SQ (2020) Online learning public opinion analysis and online teaching improvement strategies. J Commun Univ China (Sci Technol) 6:48–54+78
2. Huang FL, Feng S, Wang DL, Yu G (2017) Sentiment mining of microblog topics based on multi-feature fusion. Chinese J Comput 4:872–888

3. Li S, Zhang CM (2020) Employee public opinion monitoring and analysis in the era of big data—centering on the construction and application of network big data monitoring and analysis system. New Horizons 3:94–100
4. Ma XN, Wang T, Wang H (2018) Prediction of internet public opinion sentiment trend based on correlation vector machine. J Wuhan Univ (Nat Sci Edn) 3:237–243
5. Wang SY, Liao HT, Wu CK (2018) Topic mining of competitive enterprise news text based on sentiment classification. Data Anal Knowl Discov 3:70–78
6. Yang L, Luo WQ, Deng CL, Xiao JY (2020) Research on public opinion classification and early warning model based on grey correlation analysis. Inf Sci 8:28–34
7. Zhang SL, Wang LC (2018) Research on internet public opinion analysis technology driven by knowledge and data. J Mod Inf 4:106–111