ISSN: 1813-4890

Analysis and prediction of online public opinion based on multi-source media in the era of epidemic: A case study of Wenzhou region

Jun Pan

School of Artificial Intelligence, Wenzhou Polytechnic, Wenzhou, Zhejiang 325035, China;

526178719@qq.com

Abstract

In recent years, global outbreaks of epidemics have had a tremendous impact on people's lives and society. In the era of pandemics, public opinion analysis has become an important tool for understanding public sentiment, assessing risks, and guiding decision-making. In this study, we will utilize multiple sources of media data, including social media, news reports, online forums, etc., to conduct a network-based analysis of public opinion on the Wenzhou epidemic. We will employ text mining and sentiment analysis techniques to process and analyze large-scale textual data, aiming to reveal the changing trends in public sentiment, shifts in focus, and the impact of rumor dissemination. Additionally, we will apply social network analysis methods to explore information dissemination pathways, key nodes, and network structures in the Wenzhou epidemic, in order to gain a deeper understanding of the dynamics and mechanisms of information dissemination.

Keywords

Era of the epidemic, public opinion analysis, and multi-source media.

1. Introduction

In recent years, the global outbreak of the epidemic has had a profound impact on various fields such as society, economy, and politics. In the era of the epidemic, people are facing unprecedented challenges and changes, and online public opinion analysis has become particularly important because the epidemic has brought widespread social impact and public attention, and online public opinion has become one of the main channels for people to express and obtain information. Public opinion analysis has become one of the important tools for understanding public opinions, assessing risks, and making decisions. Especially in the era of the internet, analyzing online public opinion through multi-source media can more comprehensively and quickly capture and analyze the dynamics of social public opinion.

This article aims to explore the analysis of online public opinion based on multi-source media in the era of the epidemic, with Wenzhou as a specific case study. Wenzhou, as an important city in the southeastern coastal region of China, once became an important source of transmission of the domestic epidemic in the early stages of the epidemic. By conducting online public opinion analysis on the Wenzhou epidemic, we can gain a deeper understanding of the public's emotions, concerns, and behaviors during the epidemic period, and further understand the impact of the epidemic on society and the effectiveness of response measures.

In this study, we will use multi-source media data, including social media, news reports, online forums, etc., to conduct online public opinion analysis of the Wenzhou epidemic. We will apply Text mining and sentiment analysis technologies to process and analyze large-scale text data to reveal the changing trend of public sentiment, the shift of focus and the impact of rumor spread. The results of this study will provide important reference and guidance for governments,

ISSN: 1813-4890

organizations and decision makers. By understanding the public's needs and concerns, the government can more accurately formulate policies and measures to improve the effectiveness of responding to the epidemic. At the same time, this study also provides empirical research support for the methods and techniques of public opinion analysis, providing reference and reference for epidemic management and public opinion research in other fields.

2. Current research status at home and abroad

In recent years, scholars from both domestic and international institutions have conducted extensive research on network sentiment analysis based on multi-source media. Internationally, many studies have focused on using social media data for sentiment analysis. For example, Bastos et al. (2020) analyzed the spread of misinformation about COVID-19 worldwide using Twitter data. Zhang Yixiang et al. developed a method combining machine learning and sentiment analysis techniques to conduct network sentiment analysis on hot events in the sports field [2].Peng Yuting et al. applied a word frequency analysis method and used visualization software to visually display the high-frequency terms in the sample, conducting sentiment analysis on the overseas online public opinion of the Beijing Winter Olympics. They categorized and discussed specific topics related to "Beijing Winter Olympics" tweets, proposing strategies for risk assessment and response to overseas online public opinion on the Beijing Winter Olympics [3].Fang Yuanyuan et al. conducted empirical analysis on the COVID-19 epidemic. Firstly, they used the snowNLP library in Python to score the sentiment of Weibo comments to determine the sentiment inclination of netizens. Then, they performed descriptive analysis on the textual content of Weibo and comments using high-frequency words and word clouds. They further conducted topic extraction on the textual content and comments in different sentiment stages, combined with the actual changes in the epidemic, to explore the hot topics of netizens' concerns [4].Zhang Ting et al. (2020) analyzed public sentiment during emergencies using social media data and found that sentiment dissemination exhibits distinct spatial and temporal characteristics [5].Huang Bo et al. (2020) studied the impact of network sentiment on government decision-making using sentiment analysis and topic modeling methods. They proposed strategies and recommendations for sentiment management [6].In conclusion, the current research mainly relies on text analysis methods to study sentiment data. Most studies focus on individual data sources, and the inconsistency in data formats, as well as the complexity of data collection and processing, and subjectivity in sentiment analysis, pose challenges.

3. Project Research Content and Implementation

This article mainly divides the basic design concept into six steps, namely data collection from multiple sources, data preprocessing, international communication of hot topic words, and emotional trend analysis.

(1)Data crawling： Collect data from multiple data sources, including social media, news reports, online forums, etc.

(2)Data preprocessing: Clean and preprocess the collected data, including removing noise and processing missing data.

(3)Hot topic analysis: Extract hot topic keywords within the epidemic era by analyzing data.

(4)Sentiment intensity analysis: Using sentiment analysis technology to analyze emotional tendencies in text data and analyze changes in emotional trends.

The process is as follows:

ISSN: 1813-4890

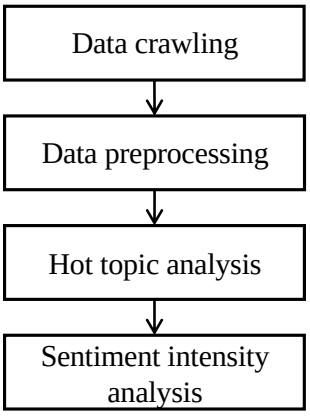

Fig.1 Implementation approach

3.1. Data crawling

The method of data collection is crucial for subsequent analysis and prediction. Currently, most methods are based on analyzing data from a single media source, which hinders objective and comprehensive analysis of the overall public sentiment. In the era of the pandemic, numerous topics related to it emerge on social media platforms, forums, and other platforms. Various unexpected topics continue to arise. Therefore, it is necessary to collect data from as many platforms as possible to make the early warning system more representative, comprehensive, and objective. Regarding data collection, this project employs a multi-source data collection method to capture real-time data from all mainstream media platforms related to the Wenzhou region on a daily basis. This includes platforms such as Weibo, WeChat Official Accounts, Toutiao, Ruian Forum, Yueqing Forum, and 703804 Forum. The collected data includes post content, comments, information of the bloggers, interaction data, and comment data. To overcome some websites' anti-scraping mechanisms, cookie verification and Selenium are used to simulate browser clicks, enabling the collection of all accessible data. The specific process is as follows:

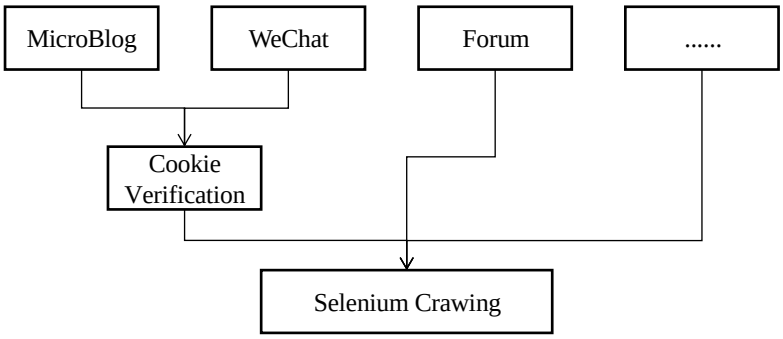

Fig.2 Data rawling

3.2. Data preprocessing

Due to various sources, multi-source data requires a lot of processing. For the captured website HTML page source code data, label parsing is performed to extract all the required content, including author information, body text, interactive data, comment data, etc. After capturing, further organization is needed to standardize the same content fields and preprocess the text data, Then remove some irrelevant content, such as useless emoticons, such as content between # and #, etc. Missing key fields and deleting duplicate fields are also necessary. Further denoising is required, as the crawled data comes from web pages and often comes with some HTML tags such as "", "<img/>", "<a>" that require denoising. For the above tags, the regular rule determined is reg=<[^>] *>, which means removing all content with "<>". Finally, we need to remove Stop word and Stop word, which refer to words with high frequency but less information and some useless symbols. The appearance of these words not

ISSN: 1813-4890

only affects the discrimination of emotional tendencies, but also serves as noise for topic extraction, so it needs to be removed.

3.3. Hot topic analysis

Due to various sources, multi-source data requires a lot of processing. For the captured website HTML page source code data, label parsing is performed to extract all the required content, including author information, body text, interactive data, comment data, etc. After capturing, further organization is needed to standardize the same content fields and preprocess the text data, Then remove some irrelevant content, such as useless emoticons, such as content between # and #, etc. Missing key fields and deleting duplicate fields are also necessary. Further denoising is required, as the crawled data comes from web pages and often comes with some HTML tags such as "", "<img/>", "<a>" that require denoising. For the above tags, the regular rule determined is reg=<[^>] *>, which means removing all content with "<>". Finally, we need to remove Stop word and Stop word, which refer to words with high frequency but less information and some useless symbols. The appearance of these words not only affects the discrimination of emotional tendencies, but also serves as noise for topic extraction, so it needs to be removed.

3.4. Hot topic analysis

During the pandemic, hot events occur from time to time, making hot topic analysis highly meaningful. The influence of hot topic propagation on social media, which spreads rapidly, should not be overlooked. As long as an event undergoes fermentation on these hot platforms and triggers intense discussions among netizens, it can appear on the trending list regardless of its scale. Influential bloggers can easily fuel the fermentation of the entire hot topic. Hot topic analysis aims to identify visually prominent "key words" that appear frequently in diverse media. If these keywords are obtained based on the content of hot topics, the more prominent the keywords, the hotter the current topic. Weibo's hot topic analysis is based on noise reduction processing, segmenting the comments into Chinese words, extracting keywords, and counting their frequency. The most frequently occurring topic in the frequency result is identified as the hot topic.

After completing data preprocessing, Chinese word segmentation is necessary, and the choice of word segmentation tool is crucial as it directly affects the final results. Word segmentation involves splitting Chinese texts into discrete words according to certain rules. The quality of word segmentation directly impacts the effectiveness of text processing. Since users often combine keywords when expressing opinions on media platforms, hot topics or sensitive topics are often reflected in certain keywords or phrases. The sentiment value of public opinion, whether "positive," "neutral," or "negative," is often found in key words as well.

Considering the large quantity of public sentiment involved in this study and the need for real-time processing, a fast word segmentation speed is required. Jieba is a powerful Chinese word segmentation library that combines dictionary-based and statistical-based methods. It has a fast word segmentation speed and comes with a built-in dictionary of over 20,000 words. It implements efficient word graph scanning and can generate all possible word combinations using Chinese characters in a sentence. Additionally, it provides the feature of importing custom dictionaries. In this case, the custom dictionary imported includes Chinese-English terms related to the COVID-19 pandemic published by the China Foreign Languages Bureau. It mainly covers disease names, policy measures, infectious control, institutional names, occupational groups, location names, medical instruments, pathological symptoms, and other medical terms. The first batch includes 180 terms, the second batch 134 terms, the third batch 95 terms, the fourth batch 104 terms, and the fifth batch 102 terms. This article combines the terms from all five batches to create a custom word segmentation dictionary for processing the original data.

ISSN: 1813-4890

The resulting word segmentation needs to be further statistically analyzed and sorted to obtain the final word frequency result. The specific process is as follows:

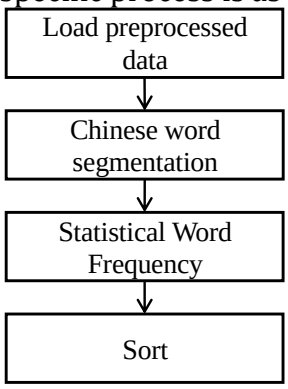

Fig.3 Hot topic analysis

3.5. Sentiment intensity analysis

From 2022 to 2023, the epidemic has relapsed again, with a long duration and a wide range of impacts, which has attracted enthusiastic attention from netizens. Therefore, the research on the evolution of public opinion in COVID-19 has important practical significance. This article first uses snowNLP to rate the emotions of each Weibo comment, and calculates the emotional rating of the day based on the weighted average of the emotional scores of all comments on that day, observing changes in public opinion. The emotional score calculated through snowNLP ranges from 0 to 1. The closer it is to 1, the more positive the emotion is. Conversely, the closer it is to 0, the more negative the emotion is. The daily emotional score obtained by weighted average of the emotional values of all media content on a daily basis. Based on the final results, we found that the average emotional fluctuation of the epidemic was basically around 0.72. It began to fluctuate positively towards negative in early December 2022, peaked around December 22, 2022, and then decreased in volatility. On March 23, 2023, it decreased to 0.61.

4. Conclusion

In summary, this study focuses on the analysis and prediction of online public opinion based on multi-source media in the era of the epidemic. Through six major steps of collecting multiple data sources, data preprocessing, analyzing hot topic words, and analyzing emotional trends, a complete analysis framework has been constructed. By collecting multi-source data, including social media, news reports, and online forums, we can obtain more comprehensive and diverse public opinion data, making the analysis results more objective and comprehensive. During the data preprocessing process, we cleaned and standardized the collected data to ensure its quality and consistency, laying the foundation for subsequent analysis. The analysis of hot topic keywords helps us identify the hot events that have emerged in the era of the epidemic, which can often spread quickly on social media and other platforms, triggering heated discussions among netizens. By analyzing the frequency and prominence of key words, we can grasp the current hot topics. Emotional trend analysis helps us reveal the changing trends of public emotions and the shift of focus. Through emotion analysis technology, we can evaluate emotional tendencies in text data and analyze them based on changes in emotional trends, in order to understand the public's emotional attitudes and evolution process towards the epidemic and related topics. However, there is still some room for improvement in the current research. The improvement in system performance includes optimization of crawling algorithms and information analysis algorithms to improve system efficiency and accuracy. In addition, with the increase of multi-source media topics, the storage space of relational databases may not meet the requirements, so in future system iterations, it may be necessary to consider using non relational databases. In summary, the analysis and prediction of online

public opinion based on multi-source media is of great significance in the era of the epidemic. Through more comprehensive and objective data collection and analysis, we can better understand the public's attitudes and emotional tendencies towards the epidemic and related events, provide valuable information and insights for decision-makers, help address the challenges brought by the epidemic, and achieve more effective public opinion management and prediction.

Acknowledgements

This article was supported by the 2022 Basic Research Project of Wenzhou Science and Technology Bureau, titled "Analysis and Early Warning of Network Public Opinion Based on Multisource Media in the Era of Epidemic: Taking Wenzhou as an Example" (Wen Ke Guan [2022] No. 11, Project No. R20220129).

References

[1] Bao, J., Luo, L.,&Wang, S. (2020) COVID-19 and online misinformation: A double edged password Journal of Medical Internet Research, 22 (8), e19512

[2] Zhang Yixiang, Zhang Xiaoli, Du Xiayu Online public opinion analysis based on the LDA topic model "Sun Yang case" [C]a//Chinese Association of Sports Sciences. Summary of the 12th National Sports Science Conference - Special Report (Sports Information Branch), 2022:96-98. DOI: 10.26914/c.cnkihy.2022.004941

[3] Peng Yuting, Wan Xiaohong Analysis of Online Public Opinion for the Beijing Winter Olympics on Overseas Social Media - Word Frequency Analysis Based on Twitter Posts [C]//Summary Compilation of Papers at the 12th National Sports Science Conference, Chinese Sports Science Society - Wall Newspaper Exchange (Sports News Communication Branch), 2022:2. DOI: 10.26914/c.cnkihy.2022.006937

[4] Fang Yuanyuan. Online public opinion analysis based on Text mining and emotion analysis [D]. Anhui University of Finance and Economics, 2021. DOI: 10.26916/d.cnki.gahcc.2021.000565

[5] Zhang Ting, Zhou Qianqing,&Zhu Hui (2020) Research on the Spatial Communication Characteristics of Public Opinion and Emotions Based on Social Media [J] Journal of Systems Engineering, 35 (5), 938-948

[6] Huang Bo, Zeng Yanlan,&Sun Kai (2020) Online public opinion analysis and policy recommendations based on Big data [J] Research on Technology Management, 40 (22), 135-142