Topic Mining and Opinion Identification of Different Disseminators in Weibo Public Opinion Propagation Cycle Post-print

Authors: Haihan Liao, Wang Yuefen, Guan Peng

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] This study investigates the public opinion hotspots and principal viewpoints concerning different disseminators within the propagation cycle of Weibo public opinion, thereby uncovering the characteristics and patterns of public opinion dissemination to furnish a foundation for public opinion analysis and decision-making. [Method/Process] Utilizing factual textual data from specific public opinion incidents as the source material, and guided by lifecycle theory and the LDA methodology, we devise a research procedure and construct a research model to conduct thematic investigations into topics from different disseminators in Weibo public opinion events. This encompasses topic extraction and semantic annotation of results, semantic analysis of topics across different disseminators at various stages, as well as identification and characterization of public opinion thematic viewpoints based on a temporal dimension. [Results/Conclusion] The research reveals that the proposed model is capable of excavating the thematic structure, opinion lineage, and features of different disseminators throughout the public opinion propagation cycle, and can identify correlated, representative, and significant words distributed within the text. Additionally, the findings demonstrate that topics in information released by official media and mass media on Weibo exhibit distinct differences from hot topics discussed by users.

Full Text

Topic Mining and Viewpoint Recognition of Different Communicators in the Microblog Public Opinion Transmission Cycle

Liao Haihan[1], **Wang Yuefen**[1,2], **Guan Peng**[1] [1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094

[2]Jiangsu Province Social Public Safety Science and Technology Collaborative Innovation Center, Nanjing 210094

Abstract

[**Purpose/Significance**] This study explores the hotspots of public opinion and main viewpoints in the communication content of different communicators during the microblog public opinion transmission cycle, aiming to discover the characteristics and patterns of public opinion transmission and provide a basis for public opinion analysis and decision-making. [**Method/Process**] Using factual text data from specific public opinion events as the source, and guided by lifecycle theory and the LDA method, this paper designs a research process and constructs a research model to conduct thematic studies on different communicators in microblog public opinion events. The research includes topic extraction and semantic annotation of results, semantic analysis of different communicators' themes at various stages, and viewpoint recognition and characterization of public opinion themes based on the time dimension. [**Result/Conclusion**] The findings indicate that the proposed research model can excavate the topic structure, viewpoint context, and characteristics of different communicators during the public opinion transmission cycle, and identify correlated, representative, and important words distributed throughout the text. Additionally, the study reveals that topics in information released by official media and mass media on microblogs differ significantly from hot topics discussed by users.

Keywords: microblog public opinion; different communicators; topic mining; viewpoint identification; lifecycle theory; LDA topic model

1 Introduction

New media has given rise to diversified forms of social public opinion transmission, instantaneous dissemination speed, massive volumes of content, high decentralization, and fragmented conversations. These characteristics often amplify or distort news reports about social events during publication and commentary processes, frequently triggering public opinion incidents. Microblogs, as a representative new media platform, not only stimulate netizens' desire for uninhibited expression through anonymity and freedom but also highlight interactions among communication elements due to convenient and social communication functions. Public opinion events posted on microblogs generate topics among thousands of discourses during transmission, with content varying between praise and criticism, diverse and complex emotions and attitudes among communicators, and multifaceted impacts. This strengthens the suddenness and effect of public opinion transmission, increases social instability factors and complexity, and intensifies the difficulty of public governance. After information is released, how do information comments generated through microblog transmission develop in terms of thematic content, and how do they compare with the original information release themes? From the perspective of real-time public

opinion monitoring, this paper combines lifecycle theory and the LDA model to construct an effective dynamic mining model for hot topics in public opinion events, tracks public opinion transmission content based on the research model, and excavates hidden topic information.

2 Related Work

2.1 Current Status of Public Opinion Theme Research

Based on research needs, this paper first reviews literature related to the current status of public opinion theme research, then summarizes the methods and theories used in the study. Through literature investigation, scholars both domestically and internationally have conducted public opinion theme research mainly from the perspectives of theme mining and theme monitoring. In theme mining research, scholars commonly apply natural language processing, text clustering, co-word analysis, topic modeling, algorithm improvement, and other technical methods. In theme monitoring research, they primarily focus on theme tracking, early warning, and other research angles.

In theme mining research, Chen Xiaomei et al. used the LDA topic model viewpoint extraction method to compare and analyze differences between viewpoint extraction methods, explained group wisdom and individual cognitive processes of network public opinion platforms from a cognitive perspective, and finally discovered the advantages and new paths of the LDA topic model for extracting public opinion viewpoints [**?**]. Zhang Shouhua et al. adopted the TFIDF method and topic clustering algorithms in public opinion hot topic research, designed a theme mining system, and implemented it through key steps including public opinion preprocessing, keyword extraction, topic clustering, and hot topic analysis. Their research found that the designed system had high recognition accuracy for network public opinion hotspots [**?**]. Li Lei et al. applied co-word analysis methods to study public opinion themes, constructed co-word matrices, identified hot topics through keyword co-occurrence, and found that the proposed method not only had practical value but could also improve the efficiency of refining and summarizing network public opinion information [**?**]. Qian Aibing applied measurement methods such as theme attention and hot topics, constructed a network public opinion analysis model based on themes including public opinion theme planning, information collection and analysis, and early warning, and obtained research conclusions on public opinion theme attention, hot topics, focal points, and key points [**?**]. Liang Xiaohe et al. constructed a hypernetwork model with four layers of sub-networks including users, viewpoints, emotions, and temporal stages through network measurement methods, and combined the model with specific case events for analysis. Their research conclusions indicated that sub-network analysis based on the hypernetwork model for public opinion theme discovery could reveal characteristic information of each sub-network, and hyperedge analysis could be used for public opinion early warning analysis, theme mining, and theme evolution analysis [**?**]. N. Li et al. conducted text mining research using K-means clustering and SVM

algorithms, categorized Sina Weibo sports forums, and mined text data to discover hot topics, finding that both methods yielded the same results [**?**]. L.Y.F. Su et al. studied public opinion themes based on the HK algorithm, improved intelligent algorithms, conducted content analysis, and mined communication theme emotions, finding that the research method had reliability and validity for social media public opinion theme mining [**?**].

In public opinion theme monitoring research, Ding Shengchun et al. adopted web crawler, webpage preprocessing, and text extraction technologies to construct a multilingual public opinion monitoring system for the South China Sea issue and implemented theme tracking under time series. Their research conclusions indicated that the constructed public opinion monitoring system could achieve public opinion information collection, processing, and analysis [**?**]. Zhang Yu et al. based on the Bass model, studied the emotional distribution of different topics within hot microblog events over time through microblog text segmentation, threshold determination, microblog text feature extraction, topic dictionary construction, and microblog text topic division. They discovered the status of different topics and the impact of group emotions on topics [**?**]. An Lu et al. adopted lifecycle theory and word2vec technology to conduct fine-grained division of topic comment emotions, calculate emotional intensity, and ultimately achieve collaborative analysis of microblog themes and emotions, finding that the proposed analysis method could reveal the collaborative evolution patterns of microblog network public opinion themes and emotional characteristics for specific events [**?**]. J. Zhao et al. applied the NB algorithm, improved the algorithm, and defined parameters to conduct emotional measurement of abnormal public opinion events, achieving the significance of public opinion monitoring under time series [**?**]. Q. Mei et al. applied the TSM model to monitor microblog themes, established a research model through parameter definition, and then applied lifecycle division to conduct empirical research on public opinion event data. The study found that the established model had good mining and monitoring effects and potential application value [**?**].

In summary, public opinion theme research is a concern for many researchers, with increasing numbers of studies forming many valuable research results and viewpoints. The above research shows that most network public opinion theme mining studies are based on text mining technology and intelligent algorithms. In theme mining, the LDA model and feature word extraction methods in natural language processing can more accurately reveal word features of corpora and are suitable for large-scale dataset mining research. Co-word analysis has uncertainty in word frequency thresholds, and social network analysis of theme words focuses on word associations, suitable for small dataset analysis. Text clustering methods have strong subjectivity in labeling categories and are more appropriate for small data studies. Most theme monitoring research is conducted from a time dynamic perspective, often involving lifecycle theory, time series, and other theories and methods. Technically, it involves word2vec technology, NB algorithm, TSM model, statistical analysis, and other methods. Lifecycle theory is typically combined with technical methods to demonstrate changes,

variations, and other forms and trends of monitoring themes. However, most current public opinion theme monitoring studies lack discrimination of different nature themes, making monitored theme content and viewpoints scattered and unfocused. Since this study will process large-scale data and aims to discover the content and viewpoints of different nature communicators' public opinion themes under time granularity, this paper adopts the LDA model method and lifecycle theory, guided by their combined application, to propose a research process and model to reveal the laws of network public opinion development.

2.2 Research Methods

2.2.1 LDA Topic Model LDA topic word mining is an important mining method in natural language processing and is also a fully generative model. The LDA topic model can display the collection and probability of related terms under a single theme, can exclude the influence of subjective factors on scientific research, and compensates for the limitation that traditional research cannot effectively mine large batches of text. Since the application of the LDA model in domestic and international research is already quite extensive and mature, this study will not elaborate further.

In public opinion theme research, Tang Xiaobo et al. mined microblog hotspots based on the LDA model, constructed an LDA model about the concept of microblog popularity, and conducted experiments through collected microblog data. The research found that the improved LDA model could obtain more intuitive microblog popularity expressions and more convincing mining conclusions [**?**]. Lin Ping et al. extracted topics based on the LDA model, analyzed topic heat changes through a post-discrete time topic model approach, and analyzed topic content migration through a pre-discrete time topic model approach. The study not only discovered public opinion event topic content but also found that the optimal number of topics is closely related to the concentration of text content focus [**?**]. W.X. Zhao et al. used LDA text mining technology to conduct comparative mining of similar themes between Twitter content and traditional media New York Times, deeply exploring the relationship between posted and replied blog themes and categories. The study also found similarities and differences between offline and online situations [**?**]. M. Pennacchiotti et al. used the LDA topic model to discover user interests and ultimately developed a system to recommend friends with similar interests to users [**?**].

In summary, the LDA topic model is not only an effective method for public opinion theme analysis but also a hot technology favored by scholars.

2.2.2 Lifecycle Theory Lifecycle theory can effectively reveal the process of things from birth, growth, maturity, decline to death, and has been widely applied in various disciplines. In network public opinion research, the entire experience process from information generation to failure is generally defined as the transmission lifecycle. As a conceptual theory, lifecycle theory requires

specific links for support, and its application has more practical significance compared to model research and simulation methods.

Research on lifecycle theory in public opinion themes includes: An Lu et al. used SOM self-organizing mapping and lifecycle theory methods to conduct comparative analysis of hot topics about the Ebola virus outbreak in West Africa on Twitter and Weibo platforms through text preprocessing and theme classification steps. The research conclusions found similarities and differences in theme evolution patterns and temporal trends [**?**]. Chen Fuji et al. adopted a topic communication evolution game model to conduct corresponding research on topic communication through lifecycle scenario prediction and model fitting, and the study also summarized countermeasures for network public opinion hot topic communication models based on evolutionary games [**?**]. Zhang Silong designed a multi-information dynamic update mechanism for topics based on "microblog attention" in his research by drawing on "microblog lifecycle" theory [**?**]. Q. Mei et al. used a new probability method to construct a public opinion theme research model and combined lifecycle theory to generate snapshots of themes for each given time period through theme lifecycle division. The research results showed that the constructed research model could be applied to analysis of general time and space information [**?**]. The characteristics of the above research show that studies guided by lifecycle theory can more deeply and meticulously mine effective information and conclusions in public opinion theme communication patterns.

3 Research Design

Based on the above description, this paper proposes a research process design for microblog public opinion theme mining and viewpoint identification based on different communicators. The design includes three components: research ideas and processes, division rules of lifecycle theory, and construction of hot topic models for different communicators.

3.1 Research Ideas and Processes

To achieve hot topic mining and viewpoint identification of different communicators during stages of the public opinion transmission cycle, the research ideas are designed as follows: First, normalize microblog text data; second, divide the stages of the public opinion event transmission cycle based on lifecycle theory; third, conduct theme mining and viewpoint identification based on the LDA model; finally, perform word frequency statistics and compare them with the aforementioned theme mining results to verify the research ideas and conclusions. The specific research process and assumptions are described below:

3.1.1 Microblog Text Data Normalization Heterogeneous data in microblogs directly affects theme extraction results. Therefore, this study must first address microblog heterogeneous data normalization, which involves cleaning all heterogeneous data types and uniformly converting them into

standardized data formats for preservation, laying the foundation for theme extraction and semantic mining. Normalization methods involve natural language processing procedures such as word segmentation, tokenization, and stop word filtering.

3.1.2 Characterization of Public Opinion Event Transmission Cycle Stages Based on Lifecycle Theory The public opinion event transmission lifecycle reflects the different development stages and vitality of public opinion information, which in turn reflects the transmission effectiveness of public opinion information content. Although in public opinion research fields, different division results may occur due to different specific cases of transmission lifecycle evolution, according to lifecycle theory, the lifecycle of general events can be divided into four stages: germination period, growth period, decline period, and stable period. This study will use these four stages as the basis for establishing rules for dividing transmission cycle stages in case studies.

3.1.3 Theme Mining and Viewpoint Identification of Different Communicators Based on the LDA Model This study conducts semantic mining research on public opinion at each window period of the transmission cycle from two levels. The first level: Conduct LDA theme extraction on the corpus according to different window periods divided by the lifecycle, and summarize feature words extracted from each stage of the public opinion event. Combine the collected corpus context to conduct mainstream phrase semantic annotation on the summarized feature words and interpret the semantic meaning of each transmission cycle stage. The second level: Based on the annotated and interpreted semantic phrases, conduct viewpoint identification of themes at each transmission cycle stage.

3.1.4 Word Frequency Statistics Verification Apply relevant tools or software to extract high-frequency words from the analysis corpus, remove noise and meaningless words, and rank them by frequency.

According to the content described in the above steps, this paper proposes a research flow design for microblog public opinion theme mining and viewpoint identification based on different communicators, as shown in Figure 1 [Figure 1: see original paper] (Figure 1 contains the specific preset research methods, tools, and key issues of each link).

3.2 Division of Microblog Transmission Cycle Stages Based on Lifecycle Theory

In the division of lifecycle theory, scholars usually divide it into three or four stages according to specific application scenarios. Here, this paper divides the microblog public opinion transmission cycle into general lifecycle stages and proposes division rules:

(1) **Germination Period**: Microblog public opinion posts and comments are relatively few, transmission increment is almost zero or even negative

growth, language is scarce and monotonous, topic types are few, but new words continue to appear, indicating that the transmission situation is in the germination period.

(2) **Growth Period**: Sina Weibo shows explosive growth with exponential growth patterns, posts and comments show growth status, and new messages and comments continue to increase. Meanwhile, the number of public opinion topics surges, and posts and comments show a surge curve over time, indicating that public opinion transmission has entered a period of rapid explosion.

(3) **Decline Period**: Public opinion event posts and comments show a rapid downward trend, transmission volume growth rate decreases, word growth rate may be negative, topic numbers basically do not update, indicating that public opinion transmission is exiting hot topics at this stage, and transmission volume has significantly decreased.

(4) **Stable Period**: After the decline period, the daily transmission volume of public opinion events enters a relatively stable period. The growth rate of transmission volume at this stage is almost zero, and after the germination, growth, and decline periods, the transmission energy has entered a period of communication group emotional venting and thought expression. However, two trends may still occur at this stage: first, the transmission volume of the theme decreases, with no new language trends, and the transmission volume maintains a certain stable level; second, on the basis of the original theme, due to disclosure of news, some word volumes show a positive growth trend, and public opinion information transmission volume increases, indicating that the event has derived new hot topics, and new public opinion is about to erupt.

3.3 Hot Topic Model for Different Communicators Based on the LDA Model

According to the divided transmission cycle, this paper constructs a hot topic research model in the microblog public opinion event environment combined with the LDA topic model. The main dimensions of the model construction include: time cycle, microblog users, microblog content, theme mining and viewpoint identification. The dimension descriptions are as follows:

(1) **Time Cycle Dimension**: The transmission of public opinion events is composed of time flow and information. According to the development trend of public opinion information transmission cycles, this study uses 1 day as the time granularity to divide public opinion into lifecycle stages.

(2) **Microblog User Dimension**: Users are the main body of public opinion release and production. This study conducts transmission theme analysis based on information sources at two levels: user posting and user commenting.

(3) **Microblog Content Dimension**: Microblog users'posting and commenting information implicitly contains users' emotional tendencies, thoughts, viewpoints, opinions, and other subjective attitudes toward public opinion events.

(4) **Theme Mining and Viewpoint Identification Dimension**: Based on information collection and LDA theme extraction, public opinion themes are obtained, and themes hidden in microblog discourse are semantically annotated. According to the semantic analysis results of themes, viewpoints are then highly summarized and extracted.

However, microblog data still has many special characteristics in the above dimensions: At the time cycle dimension level, themes held by public opinion events often change within a short time, meaning multiple themes appear within the same stage. At the user dimension level, the same user may post different content in different periods or post similar content. At the content dimension level, one microblog post may express several viewpoints, or several microblog posts may belong to the same viewpoint. Therefore, based on the characteristics and particularities of microblog data, this study constructs a theme mining model under the fusion influence of three dimensional factors: time cycle, microblog users, and content. The theme mining and viewpoint identification dimension is regarded as the result of the action of these three factors. The research model is shown in Figure 2 [Figure 2: see original paper]. This paper strives to construct a hot topic research model for different communicators that can reflect the situation of public opinion transmission and provide valuable conclusions for real-time monitoring and crisis response of public opinion management and control.

4 Empirical Analysis

4.1 Research Data Collection and Processing

4.1.1 Data Collection and Basic Description This study used the web crawler tool GoSeeker to collect data on the "8.12 Tianjin Explosion Incident." The collected data includes: microblog post content, microblog comment content, microblog poster ID, microblog commenter ID, microblog ID, etc. The collection period was from August 12, 2015, to September 13, 2015. Figure 3 [Figure 3: see original paper] shows the statistical trend of the total daily transmission volume of collected microblog posts and comment data.

According to the transmission volume trend of the "8.12 Tianjin Explosion Incident," this paper divides the public opinion transmission cycle of the case study into three stages: germination period (August 12 to August 15), outbreak period (August 16 to August 31), and stable period (September 1 to September 13). In the division of stages, the transmission volume surged on August 16 and reached the highest peak, which is a significant characteristic of the growth period; from August 17 to 31, the total transmission volume showed a rapid decline trend, which is a significant characteristic of the decline period. Since

both transmission situations exist in the high-temperature fermentation period of the event, if we conduct separate measurement and analysis by stage, it would be detrimental to theme mining. Therefore, these two stages are combined and defined as the outbreak stage. The transmission trends of the remaining stages conform to the division rules of the public opinion lifecycle.

4.1.2 Text Processing Experimental Steps According to the different natures of microblog information release, this study defines communicators as microblog posters and microblog commenters. Microblog posters include official media, mass media, and other public media, while microblog commenters include most ordinary users and other information receivers. For different types of communicators, the study establishes two types of corpora: a microblog poster corpus and a microblog commenter corpus. According to research needs, this paper selects long blog posts posted by microblog users and short comments by users as the initial corpus. Since the collected microblog texts of the "8.12 Tianjin Explosion" incident contain rampant black words, vulgar language, fragmented discourse, and rumors, with intense, obscure, and highly irregular writing forms, the initial corpus text language is chaotic and disorderly. Considering the efficiency of text analysis, this paper summarizes and categorizes the viewpoints and semantics expressed in microblogs in combination with blog text content, establishes a standardized basic corpus after manual noise processing.

The study then uses the jieba word segmentation toolkit to implement Chinese word segmentation, remove stop words, and other natural language processing, eliminating meaningless words such as "有点" (a bit), "木子" (a name), "感觉" (feel), "无法" (unable), "呵呵" (hehe), to obtain neatly structured data as the corpus for experiments. Then, based on the open-source gensim package, LDA topic model parameter training is implemented. The LDA parameter settings in this study refer to relevant literature methods [**?**], with iteration times set to 2000, hyperparameters set as α=0.01, β=0.05, k=10. After determining parameter values, the corpus file is input, and the LDA modeling program is run. After LDA theme extraction, two important result documents are obtained. One is the theme distribution document, used to calculate theme intensity; the other is the feature word distribution document, which contains the terms and probabilities of feature distribution under each theme.

4.1.3 Theme Extraction Results Display The study conducted LDA theme extraction on the two experimental corpora for the germination period, outbreak period, and stable period, obtaining 10 extracted themes and related terms under each theme for each stage. Due to space limitations, this paper selects partial theme documents as result displays, shown in Table 1 and Table 2 .

To more accurately mine and interpret theme semantics, this study selects the top three hot themes by intensity at each transmission cycle stage for analysis, and preferably selects ten feature words with higher probabilities under hot themes for theme interpretation. The hot theme feature word induction results are shown in Table 3 .

Based on Table 3, this paper combines the corpus to conduct mainstream semantic annotation on feature words to obtain keyword groups of hot themes at each stage, with results shown in Table 4 .

4.2 Semantic Analysis of Different Communicators' Themes in the Microblog Public Opinion Transmission Cycle

(1) **Interpretation of Different Communicators' Language Features**. Observing the overall content of Table 4, we find that microblog posters' language is more standardized, professional, and uses more declarative nouns, such as "dangerous goods warehouse," "safety distance regulations," "file for investigation," "fundamental governance system," etc. Microblog commenters' language is more colloquial, everyday, and uses more verbs, such as "write comments," "brush news," "see mushroom cloud," etc.

(2) **Semantic Analysis of Different Communicators' Themes at Each Transmission Cycle Stage**. In terms of microblog poster theme mining: In the germination period, the hot theme posted by microblog posters is "accident reporting," which includes main information related to the hot accident, such as containers, dangerous goods warehouses, etc., and key figures related to the accident, such as enterprise responsible persons and injured personnel. In the outbreak period, microblog posters' topics have shifted compared with the previous stage. The original release and reporting of accident-related information have transformed into "accident investigation." The new hot theme reflects the posters' in-depth mining of accident information. From the logic composed of keyword groups under this theme, we can see that the information discloses the causes of the enterprise explosion, such as violation of chemical industry standards, and transmits news about accident accountability, investigation, and filing. In the stable period, the hot theme posted by microblog posters is "post-disaster inventory." From the keyword groups, we can see that the content involves economic compensation, loss inventory, and governance requirements for enterprise management problems.

In terms of microblog commenter theme mining: In the germination period, the hot theme of microblog user comments is "accident discussion." From this theme, we can understand that at the germination stage, users receive news by brushing news and watching videos; users also discuss medical conditions, smoke situations, and call for positive energy speech. In the outbreak period, the hot theme of microblog user comments is "accident emotion." Words such as "pay for," "too heartbreaking," "give an account," and "pray" under this theme all represent users'emotions of grief, anger, helplessness, and psychological state of pursuing truth during this time period. In the stable period, the hot theme of microblog user comments is "post-disaster arrangement." The appearance of words such as "trust the government," "protect the people," and "safety supervision"in the keyword groups of this theme indicates that the public

believes the government will introduce effective measures and policies to arrange people' s livelihood issues after the disaster, and shows the psychological appeal for legal sanctions against those responsible for the accident.

From the above analysis, we can see that the hot themes of microblog posters and commenters in the germination period are "accident reporting"and "accident discussion," respectively; in the outbreak period, they are "accident investigation" and "accident emotion," respectively; and in the stable period, they are "post-disaster inventory" and "post-disaster arrangement," respectively. Comparing themes across stages, we find that content released by government officials and news media basically belongs to event information disclosure, while content posted by users basically expresses viewpoints, emotions, and psychological appeals. Moreover, there are significant differences between microblog poster themes and commenter themes.

4.3 Hot Topic Viewpoint Identification of Different Communicators Based on Theme Mining Results

Based on the theme mining results, combined with specific contexts of microblog corpora, this paper selects representative microblogs and further conducts manual identification and summary of theme viewpoints. Figure 4 [Figure 4: see original paper] shows the hot topic viewpoints and transmission evolution situation organized in this study for each stage.

Based on Figure 4, we find that during event transmission, the topic viewpoints of microblog posters and microblog commenters each have characteristics and biases. At the same time, Figure 4 also shows the structure and context of the transmission themes of the public opinion event in this study. For example, microblog posting discusses "accident reporting," "accident investigation," and "post-disaster inventory," while user commenting discusses "event discussion," "emotional expression," and "post-disaster arrangement." From the theme evolution context of the transmission cycle, we can see that the hot topics mined based on cases conform to the logic of event development and the speaking logic of posting and commenting objects. This demonstrates that the public opinion event hot topic analysis system combining lifecycle theory and the LDA model proposed in this paper has scientificity and effectiveness.

4.4 Validation of Research Results

To verify the effectiveness of the theme mining conclusions and the reliability of the theme analysis system based on the combination of LDA and lifecycle theory, this study used word frequency order analysis to count the Top 10 high-frequency words and word frequencies of public opinion content in the "8.12 Tianjin Explosion" event transmission cycle. The statistical results are shown in Table 5 .

From Table 5, we can see that the Top 10 high-frequency keywords of microblog public opinion transmission cycle represent 10 research themes globally and at

each stage, meaning that research themes are defined from single keywords. For example, in the global high-frequency keywords, the keyword "explosion"appears 417 times, indicating that "explosion" is the most concerned theme, but the deep semantic information contained under this theme cannot be known. In contrast, the result obtained by the LDA method is a theme and related terms under that theme. Comparing Table 5 with the extracted theme keyword results, from the global high-frequency word column in Table 5, we can see that the theme words extracted by microblog posters and commenters basically cover high-frequency words, indicating that the terms extracted by LDA have accuracy.

Observing and comparing the results of high-frequency keywords at each stage through the same method, such as the third column in Table 5 representing high-frequency words of microblog posters in the germination stage, and comparing them with the results of theme extraction for microblog posters in the germination period in Table 1, we can see that LDA feature words also basically cover high-frequency terms. The fifth column of keywords in Table 5 shows high-frequency words such as Tianjin, enterprise, new district, Tianjin City, and houses. The study finds that the semantic contexts between these high-frequency words span greatly, the relationships between terms are unknown, and arranging them according to frequency makes it difficult to identify semantic information. In contrast, the theme extraction results obtained by LDA, such as the scene, personnel, accident, location, explosion, logistics, warehouse, and port words displayed in the germination stage of microblog posters in Table 3, are feature words included under the "accident reporting" theme. The collection of these terms also reflects the meaning of that theme. Moreover, from the LDA feature terms, we can see that the information readable from LDA extraction results is not only specific but also more rich and diverse.

5 Conclusion

This study takes the "8.12 Tianjin Explosion" incident on Weibo as an example to conduct theme mining and viewpoint identification analysis combining the LDA model and lifecycle theory for content posted by different communicators. The research draws some meaningful conclusions, summarized as follows:

In terms of theoretical significance, this paper combines the LDA model and lifecycle theory, achieving the fusion application of methodology and theory. By applying the constructed LDA hot topic mining model to specific event cases, through dimension analysis, level analysis, and role analysis of the mining model, it identifies correlated, representative, and important words distributed in transmission, as well as the similarities and differences of hot topics among different communicators, greatly improving the interpretability of corpus information. However, in terms of overall theme analysis, topics from government official media and mass media differ significantly from user group topics, obviously showing that the Weibo platform has nurtured personalized topics among public users. Meanwhile, lifecycle theory, aimed at depicting transmission themes, achieves the function of showing macro situation structures. From a micro perspective,

this paper mines theme content and related information under time granularity, displays influential topics at each stage, deepens social public opinion research, and provides more information for decision-making. In terms of practical significance, this paper has real-time monitoring significance. The identified theme viewpoints can understand public opinion situation changes, evolution processes, and transformation of public opinion concepts, and the division observation with the help of the lifecycle serves the purpose of public opinion monitoring.

However, this study still has certain limitations and difficulties. The research model established in this paper can effectively mine hot topics in the public opinion transmission cycle, but the method still needs improvement. LDA theme model extracted feature words cannot completely interpret the meaning of a sentence like humans can, and the conclusions that can be mined are limited to the meaning of themes expressed under concentrated term clustering, which is also the limitation of LDA methods in interpreting text. Therefore, how to mine more effective information in speech remains a problem that needs further resolution. In terms of future research directions, this paper considers the correlations between public opinion transmission themes and themes, themes and events, themes and media, application scenarios, and other elements. In terms of research methods, it strengthens and deepens qualitative analysis to make the research more social scientific. It also strives to make future research more deeply depict the inheritance and variation of hidden themes in public opinion events and more clearly show the context and trends of event development through these new perspectives, elements, and methods. How to conduct semantic mining of public opinion events from a theme correlation perspective will be the focus of future research.

References

[1] Chen Xiaomei, Gao Cheng, Guan Xinhui. LDA Topic Model Method for Network Public Opinion Viewpoint Extraction[J]. Library and Information Service, 2015, 59(21): 21-26.

[2] Zhang Shouhua, Liu Zhenpeng. Research on Network Public Opinion Hot Topic Clustering Method[J]. Small Microcomputer System, 2013, 34(3): 471-474.

[3] Li Lei, Liu Ji, Zhang Hongkui. Research on Network Public Opinion Topic Discovery and Situation Evolution Based on Co-occurrence Analysis[J]. Information Science, 2016, 34(1): 44-47, 57.

[4] Qian Aibing. Network Public Opinion Analysis Model Based on Themes and Its Implementation[J]. Modern Library and Information Technology, 2008(4): 49-55.

[5] Liang Xiaohe, Tian Ruya, Wu Lei, et al. Microblog Public Opinion Theme Mining Method Based on Hypernetwork[J]. Information Studies: Theory & Application, 2017, 40(10): 100-105.

[6] LI N, WU D D. Using text mining and sentiment analysis for online forums hotspot detection and forecast[J]. Decision support systems, 2010, 48(2): 354-368.

[7] SU L Y F, CACCIATORE M A, LIANG X, et al. Analyzing public sentiments online: combining human- and computer-based content analysis[J]. Information, communication & society, 2017, 20(3): 406-427.

[8] Ding Shengchun, Gong Silan, Zhou Wenjie, et al. Real-time Monitoring Research of South China Sea Public Opinion Based on Knowledge Base and Theme Crawler[J]. Journal of Intelligence, 2016, 35(5): 32-37.

[9] Zhang Yu, Li Bing, Liu Chenyue. Research on Topic-Oriented Microblog Hot Topic Public Opinion Monitoring—Taking the "Beijing Odd-Even Number Limit Regularization" Public Opinion Analysis as an Example[J]. Chinese Information Journal, 2015, 29(5): 143-151, 159.

[10] An Lu, Wu Lin. Evolution Analysis of Emergency Event Microblog Public Opinion Integrating Theme and Emotional Features[J]. Library and Information Service, 2017, 61(15): 120-129.

[11] ZHAO J, DONG L, WU J, et al. Moodlens: an emoticon-based sentiment analysis system for chinese tweets[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012: 1528-1531.

[12] MEI Q, LING X, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM, 2007: 171-180.

[13] Guan Peng, Wang Yuefen. Analysis of Author Research Interest Evolution in Discipline Field Lifecycle[J]. Library and Information Service, 2016, 60(19): 116-124.

[14] Tang Xiaobo, Xiang Kun. Hotspot Mining Based on LDA Model and Microblog Popularity[J]. Library and Information Service, 2014, 58(5): 58-63.

[15] Lin Ping, Huang Weidong. Network Public Opinion Event Topic Evolution Analysis Based on LDA Model[J]. Journal of Intelligence, 2013, 32(12): 26-30.

[16] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[C]//European conference on information retrieval. Berlin, Heidelberg: Springer-Verlag, 2011: 338-349.

[17] PENNACCHIOTTI M, GURUMURTHY S. Investigating topic models for social media user recommendation[C]//Proceedings of the 20th international conference companion on World wide web. New York: ACM, 2011: 101-102.

[18] An Lu, Du Tingyao, Yu Chuanming, et al. Evolution Patterns and Temporal Trends of Microblog Themes in Public Health Emergencies—Taking Ebola Microblogs on Twitter and Weibo as Examples[J]. Information and Documentation Services, 2016(5): 44-52.

[19] Chen Fuji, Huang Jiangling. Research on Network Public Opinion Hot Topic Transmission Model Based on Evolutionary Game[J]. Information Science, 2015, 33(11): 74-78.

[20] Zhang Silong. Research on Microblog Hot Topic Prediction Technology[D]. Zhengzhou: PLA Information Engineering University, 2013.

[21] MEI Q, LIU C, SU H, et al. A probabilistic approach to spatio-temporal theme pattern mining on weblogs[C]//Proceedings of the 15th international conference on World Wide Web. New York: ACM, 2006: 533-542.

[22] Guan Peng, Wang Yuefen. Scientific Literature Theme Mining Based on LDA Topic Model and Lifecycle Theory[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(3): 286-299.

Author Contributions

Liao Haihan: Proposed research ideas, organized and analyzed data, wrote the paper; Wang Yuefen: Expanded research ideas, finalized the paper; Guan Peng: Data collection and processing.

Note: Figure translations are in progress. See original paper for figures.