Construction and Empirical Analysis of an Online Public Opinion Evolution Index: Postprint

Authors: Huang Wei, Zhu Zhenyuan, Xu Yejing, Sun Yue

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] This paper proposes and constructs an online public opinion evolution index to characterize the phenomenon where new sub-topics frequently emerge during the evolution of online public opinion, which holds important theoretical and practical significance for public opinion early warning and prediction. [Method/Process] Based on text clustering results and text clustering validity, discrimination criteria for online public opinion evolution and the construction process of the public opinion evolution index are proposed, and empirical analysis is conducted using the "textbook deadbeat" incident as sample data. [Results/Conclusions] The constructed public opinion evolution rate index can be used to describe public opinion evolution. During the outbreak stage, the topic public opinion evolution index is highest, thereafter gradually declining; this stage witnesses the most intense public opinion evolution with explosive growth in sub-topics. The public opinion evolution index exhibits a stepped decline during the spread period and thereafter remains negative, as sub-topics of public opinion begin to gradually decrease and the public opinion content itself shifts from divergence to convergence. After entering the dissipation period, the number of sub-topics stabilizes. As a measure of public opinion evolution rate and a discrimination method for public opinion evolution, the public opinion evolution index provides a novel perspective for public opinion monitoring and early warning.

Full Text

Construction and Empirical Analysis of Network Public Opinion Derivative Index

Authors: Huang Wei, Zhu Zhenyuan, Xu Yejing, Sun Yue
Affiliation: School of Management, Jilin University, Changchun 130022

Abstract

[**Purpose/Significance**] This paper proposes and constructs a network public opinion derivative index to describe the phenomenon where new sub-topics frequently emerge during the evolution of network public opinion, which holds important theoretical and practical significance for public opinion early warning and prediction. [**Method/Process**] Based on text clustering results and text clustering validity, this study proposes discriminant standards for network public opinion derivation and the construction process of a public opinion derivative index, using the "Classic Deadbeat" incident as sample data for empirical analysis. [**Result/Conclusion**] The constructed public opinion derivative rate index can effectively describe public opinion derivation. During the outbreak phase, the topic derivative index reaches its highest value and subsequently declines gradually. This phase witnesses the most intense public opinion derivation, with sub-topics showing explosive growth. The derivative index exhibits a stepwise decline during the spread phase, after which it remains negative, indicating that sub-topics begin to decrease gradually and the content of public opinion itself shifts from divergence to convergence. After entering the dissipation phase, the number of sub-topics stabilizes. As a measure of public opinion derivation rate and a discriminant method for public opinion derivation, the derivative index provides a completely new perspective for public opinion supervision and early warning.

Introduction

Compared with traditional media, network public opinion features faster dissemination speed, larger information volume, stronger interactivity, and lower entry barriers [1]. In the online environment, social individuals can publish and obtain information more conveniently. The accompanying high uncertainty of public opinion—the probability of new sub-topics emerging from public opinion grows exponentially, and public opinion derivation exhibits high time sensitivity—poses challenges to public opinion supervision on government new media platforms in two aspects: first, similar public opinion events produce linkage effects, touching the nerves of public opinion audiences from multiple angles and creating the possibility of public opinion eruption; second, compared with original topics, sub-topics undergo irreversible changes, producing "populist" feedback different from the connotation of the original topic, such as reports like "Ping An Executive is a Deadbeat" and "Tangshan Court Inaction." Such topic trends are difficult to predict, and there is a risk of negative polarization in the viewpoints of public opinion audiences. Government departments and public opinion-related management agencies should adopt a highly prudent attitude toward the derivative changes of hot public opinion to better achieve the national goal of "creating a clean cyberspace." Meanwhile, extracting the phenomenon of network public opinion derivation from the overall evolution process of network public opinion and constructing a separate index holds important theoretical and practical significance for real-time monitoring and rapid

response to network public opinion derivation.

Currently, relevant research on network public opinion derivative indexes both domestically and internationally mainly focuses on two aspects:

(1) Treating public opinion derivation as a new element added to public opinion dissemination models to study public opinion dissemination patterns. Based on different characteristics of network derivative public opinion dissemination, Gao Bin et al. divided network derivative public opinion into five types and proposed corresponding probability calculation methods for network derivative public opinion models, describing the specific analysis process for network derivative public opinion [2]. K. Saito et al. analyzed various node attributes in network public opinion dissemination to derive the probability of public opinion dissemination at each node [3]. D. J. Watts et al.'s research indicates that whether the starting point of public opinion dissemination is an opinion leader is not a key factor in public opinion evolution [4]. Lan Yuexin et al. built a mathematical model for network public opinion derivative effects based on the logistic model, studying the dissemination characteristics of positive and negative derivative public opinion under different information alienation conditions through model equilibrium points and stability analysis [5]. Chen Fuji et al. introduced topic derivation rate into the infectious disease dissemination model, constructing a SEIRS network public opinion dissemination evolution model. By solving the dissemination threshold and equilibrium points, they theoretically analyzed the impact of topic derivation rate on dissemination trends, and analyzed the influence of different factors on network public opinion dissemination patterns through numerical simulation experiments [6]. Yin Xicheng et al. studied the process of derivative topics and original topics disseminating independently and influencing each other in networks, concluding that derivative topics would create new peak points in the public opinion dissemination process, significantly increase topic forwarding rates, and extend the relaxation time of topic evolution [7]. Wang Lijun et al. summarized three different derivative chain structures based on the common patterns of network public opinion derivation and provided corresponding derivative probability algorithms [8]. Overall, this type of research focuses on studying the impact of topic derivation and evolution on the overall evolution and dissemination process of public opinion, without providing specific definitions for discriminant standards of public opinion derivation.

(2) Establishing relevant indicator systems for specific public opinion events. Jin Xiaohong et al. combined previous research on indicator system construction to build a thematic event public opinion indicator system from five dimensions, and using food safety as an example, obtained indicator weights through the analytic hierarchy process to construct a public opinion index [9]. He Enfeng et al. explored the potential influence of public opinion from aspects of dissemination media, scope, speed, emotional ten-

dency, and relevance, also using the analytic hierarchy process to obtain weight coefficients for each factor in the public opinion potential influence indicator system [10]. Deng Shangmin et al., based on AHP and survey methods, designed corresponding warning source indicators and warning symptom indicators for university network public opinion safety assessment to construct a university network public opinion safety assessment indicator system [11]. This type of research mainly focuses on constructing comprehensive systems for describing public opinion evolution or public opinion early warning and the weights of various indicators in the system, without establishing separate discriminant standards and corresponding indicators for the phenomenon of public opinion derivation.

In summary, current domestic and international research on public opinion derivation mostly incorporates public opinion derivation as a factor into the study of public opinion dissemination processes. There remains research space for quantitative discrimination of public opinion derivation itself. Various public opinion-related indicator studies have not established separate indicators for the phenomenon of network public opinion derivation. In addition to academic research, there are also relevant commercial applications for public opinion or network hotspot index research in the industry. Globally, the network public opinion derivative index studied in this paper has certain similarities with Google Trends developed by Google. Google Trends can provide topics related to user-input keywords based on the correlation of user search keywords, but its underlying algorithm differs from the research approach of text clustering and text clustering validity adopted in this paper. Moreover, the related topics it provides are not necessarily all sub-topics of the keyword. If the scope is limited to Chinese, the most widely used network index in the industry currently is Baidu Index, which focuses on keyword search trends and user profile construction, without providing a clear representation of the derivation and extinction process of sub-topics.

2. Theoretical Framework of Public Opinion Derivation Coefficient

2.1 Public Opinion Evolution and Public Opinion Derivation

Currently, there is no unified definition of "public opinion evolution" in research on public opinion evolution, and concepts such as "evolution," "development," and "change" in many studies have no essential difference [12]. Combining previous research on network public opinion evolution, this paper distinguishes the concepts of network public opinion "evolution" and "derivation" as follows:

(1) Network public opinion evolution refers to the overall process of a single network public opinion from occurrence, development, peak, fluctuation to fading and extinction across multiple dimensions including time, space, scale, issue, heat, and audience groups. Under this concept, domestic scholars have used different mathematical models to elaborate on

how specific public opinion events evolve. Zhou Xin et al. revealed how multimedia technology affects network public opinion dissemination patterns, supported by multimedia technology, public opinion analysis theory, and information dissemination theory, providing in-depth analysis of traditional network public opinion dissemination models [13]. Huang Wei et al. constructed a Weibo public opinion information aging model to provide computational support for monitoring Weibo public opinion information [14].

(2) In the research scope of this paper, network public opinion derivation specifically refers to the process where new sub-topics and sub-public opinions are derived during the evolution of a single network public opinion. As mentioned in the introduction, current research on public opinion derivation mostly considers public opinion derivation as a single element incorporated into public opinion dissemination models, lacking in-depth exploration of specific standards for public opinion derivation.

2.2 Network Public Opinion Derivative Index

Based on the previous definition of network public opinion derivation, this paper defines the Network Public Opinion Derivative Index (PODI) as follows: The network public opinion derivative index refers to the rate at which a single network public opinion derives new sub-topics at a specific moment. Public opinion derivation is a process where the richness of content and complexity of public opinion text change during the evolution of public opinion. As mentioned in the introduction, existing network public opinion indicator systems have not constructed separate indicators for the phenomenon of network public opinion derivation. The network public opinion derivative index obtained through text clustering and clustering validity provides a quantitative discriminant standard for public opinion derivation. Compared with previous indicator systems focusing on the overall process of network public opinion evolution, this index focuses on the single element of public opinion derivation. Relevant departments can achieve public opinion derivation early warning by monitoring the network public opinion derivative index, thereby conducting targeted public opinion control.

3. Construction of Network Public Opinion Derivative Index

3.1 Construction Process of Network Public Opinion Derivative Index

The construction process of the network public opinion derivative index is as follows: After obtaining network public opinion data, data preprocessing is required first. This process includes standardizing non-standardized data, segmenting standardized pure text public opinion data, removing stop words, and constructing a public opinion bag-of-words space. Second, TF-IDF values are used to calculate specific word weights, and then K-Means clustering is applied to text clustering results under different initial K values. Finally, clustering

validity results of text clustering under different K values are compared to determine the optimal number of topics K at the current moment as the discriminant standard for public opinion derivation. As shown in [Figure 1: see original paper]:

3.2 Text Preprocessing and Bag-of-Words Space Construction

To conduct text clustering on network public opinion, public opinion data must first be preprocessed to transform non-standardized public opinion data into standardized pure text data. Subsequently, text segmentation is required. This paper uses the jieba segmentation toolkit in the Python environment to achieve Chinese text segmentation, employing its precise segmentation mode to most accurately separate sentences in the text to meet text analysis needs. After segmentation, stop words such as function words and pronouns are removed to improve the information density of the corpus. Although the jieba toolkit already includes stop word removal functionality, it is mainly designed for its own text analysis tools and is not conducive to subsequent analysis processes. Therefore, this paper separately adopts a stop word list containing 1,893 common Chinese stop words for stop word removal. After removing stop words, all document word sets are counted, and vectors are constructed for each document, with vector values representing the frequency of a particular word in that document. Thus, the bag-of-words space VSM (vector space model) is successfully constructed.

3.3 Calculating Specific Word Weights Using TF-IDF

For the constructed bag-of-words space, this paper adopts the TF-IDF (term frequency-inverse document frequency) method to transform word occurrence frequencies into weights in the corpus. This method holds that the importance of a word is directly proportional to its frequency in a single text and inversely proportional to its frequency in the corpus.

In this method, term frequency (TF) refers to the frequency of a given word in a file, which is a normalization of term count. The specific expression is as follows:

tfi,j = ni,j / Σk nk,j

where ni,j is the occurrence count of word ti in document dj, and the denominator is the sum of all word occurrences in document dj.

Inverse document frequency (IDF) is a measure of the general importance of a word in the corpus. For a specific word's IDF, it can be obtained by dividing the total number of documents by the number of documents containing the word and taking the logarithm of the result:

idfi = log(|D| / (1 + |{j: ti dj}|))

where |D| is the total number of documents in the corpus, and |{j: ti dj}| is

the number of documents containing word ti. Of course, if the word is not in the corpus, it would cause division by zero, so generally 1 + |{j: ti dj}| is used.

The final TF-IDF value is:

tfidfi,j = tfi,j × idfi

This value is a comprehensive weight obtained after considering both high term frequency in a specific document and low term frequency in the entire corpus, thus tending to filter out common words while retaining important high-information words.

By calculating the matrix after the bag-of-words space vectors, its columns represent the collection of all document words, each row represents a document, and the vector values represent the weight of that word in both the overall corpus and that specific text.

3.4 Text Clustering Based on K-Means Algorithm

Using the matrix calculated with TF-IDF values, multiple methods can be employed for clustering analysis. In this study, we use the K-Means algorithm for text clustering. The K-Means algorithm is a classic partition-based clustering algorithm. Its basic principle is to first randomly select K documents (after TF-IDF calculation, these are vectors in the matrix) as initial cluster centers, then assign remaining documents to the most similar clusters based on the average value of objects in the cluster, while simultaneously updating the cluster averages. This process is repeated iteratively for a certain number of times until cluster division no longer changes. The specific calculation steps are as follows:

(1) Input corpus and randomly select k rows as initial cluster centers.
(2) Assign each remaining data point to the cluster with the nearest Euclidean distance ($Dis2(x,y) = \sqrt{(\Sigma(xi - yi)^2)}$).
(3) Update cluster sets C and cluster means.
(4) Repeat the above process until the objective function (Σ argmin$\|xi - cj\|^2$) converges.

The K-Means algorithm is widely used in text clustering due to its simplicity and efficiency.

3.5 Optimal Topic Number Determination Based on Text Clustering Validity Index

In traditional K-Means clustering, besides the selection of initial cluster centers, the selection of K value itself is also crucial. Generally, K value selection should be based on industry experience, with no clear theoretical guidance. Most scholars use the empirical rule of $kmax < \sqrt{n}$. Based on the assumption that "the number of sub-topics contained in a single public opinion topic should not exceed 20," this paper selects K in the range [2, 20] for 19 text clusterings, and

compares clustering validity coefficients under different K values to obtain the current optimal topic number K .

According to Zhou Kaile et al.'s research [15], clustering validity indicators can be divided into three categories: internal validity indicators, external validity indicators, and relative validity indicators. Since text clustering for network public opinion is an unsupervised learning process, external information is unavailable, and internal validity indicators are the most widely used clustering validity indicators. For internal indicators, they are usually divided into three types: indicators based on dataset fuzzy partition, indicators based on dataset geometric structure, and indicators based on dataset statistical information. Indicators based on dataset sample geometric structure evaluate clustering results according to the statistical characteristics of the dataset itself and clustering results, and select the optimal cluster number based on clustering quality. This type of indicator includes the Davies-Bouldin (DB) indicator, Calinski-Harabasz (CH) indicator, Dunn indicator, etc. This paper adopts the most commonly used DB indicator as the basis for judging text clustering validity.

The DB indicator uses the distance from intra-class samples to their cluster centers to calculate intra-class compactness, and uses the distance between cluster centers of different clusters to represent inter-class separability. It is specifically defined as:

DB(k) = (1/k) Σi=1^k maxj=1~k, j≠i (Wi + Wj) / Cij

where K is the number of clusters, Wi represents the average distance of all samples in class Ci to their cluster center, Wj represents the average distance of all samples in class Ci to the cluster center of another class Cj, and Cij represents the distance between the cluster centers of classes Ci and Cj. According to the definition of the DB indicator, the smaller the indicator value, the lower the similarity between classes, thus corresponding to better clustering results.

By calculating the DB indicator under 19 different K values, the K value with the minimum DB indicator is selected as the optimal topic number K . When K changes, it can be considered that new sub-topics have emerged in the original public opinion data (K increases), or some sub-topics have disappeared (K decreases).

3.6 Network Public Opinion Derivative Index Combined with Time Series

Based on the optimal topic number K at different moments, combined with the time series of the public opinion derivation process itself, the network public opinion derivative index PODI (public opinion derivative index) at a specific moment Ti can be obtained. It is specifically defined as:

PODI = (Ki - Kc) / (Ti - Tc)

where Ki is the optimal topic number at moment Ti, Kc represents the K

before the optimal topic number changed to the current optimal topic number, and Tc represents the moment when the optimal topic number became Ki . According to the definition of PODI, since K shows a stepwise change trend, PODI also shows a stepwise change trend, and when K does not change, it shows a downward trend over time.

4. Empirical Analysis of Network Public Opinion Derivative Index

4.1 Data Source Selection

This paper uses the "Classic Deadbeat" incident that sparked heated discussion in cyberspace between November and December 2017 to conduct empirical research on the network public opinion derivative index. On November 22, 2017, because the defendant Huang Shufen claimed she had no money and refused to compensate the plaintiff Zhao Xiangbin 850,000 yuan according to the judgment of Tangshan Fengrun District People's Court on June 8, 2017, the plaintiff Zhao Yong posted a Weibo titled "See What a Classic Deadbeat Is" and exposed his conversation with Huang Shufen when urging her to fulfill the legal judgment, causing a huge reaction in the national cyberspace. This is a typical current network public opinion incident, with related public opinion volume reaching its peak on November 23, 2017, and gradually aging thereafter. Based on this, this paper selects relevant information about this incident on three self-media platforms—Weibo, WeChat Official Accounts, and Toutiao—as data sources, searching with three keywords: "Classic Deadbeat," "Huang Shufen," and "Serious Mr. Zhao."

4.2 Data Collection

The time window for data collection in this paper is from 0:00 on November 22, 2017, to 0:00 on December 12, 2017. A total of 26,848 pieces of information (including original posts, forwards, and comments) about the "Classic Deadbeat" incident were collected from WeChat Official Accounts, Weibo, and Toutiao. Database fields include username, user ID (UID), title, author, publication time, publication content, crawl time, picture tags, picture content, video address, video description, etc. The data collection process for the "Classic Deadbeat" incident on the three self-media platforms is as follows: For Weibo data, the "Octopus" data collector was used to collect data after searching with the three keywords on Weibo. The collection content included all original posts, forwards, and comments on Weibo, publication time and user information, contained picture descriptions, contained video description information, etc. A total of 14,652 relevant original Weibo posts, forwards, and comments were collected. For WeChat Official Accounts, the "Search" function built into WeChat was used to search the three keywords, and the displayed official account content was manually collected. The collection content included official account name, official account description information, number of original articles, of-

ficial account article titles, article authors, publication time, article content, article reading quantity, article like quantity, contained picture descriptions, contained video descriptions, video addresses, comment time, comment username, comment content, etc. A total of 8,554 relevant WeChat official account articles and comments were collected. For Toutiao, the web version of Toutiao was searched with the three keywords and manually collected. The collection content included article title, author, article content, publication time, picture description, comment user, comment content, comment publication time, etc. A total of 3,642 relevant Toutiao articles and comments were collected. Other descriptions of the collected data are shown in :

Table 1 Description of Collected Data

Platform	Data Volume (pieces)	Average Characters per Piece	Image/Video Quantity (pieces)	Average Comments per Piece
Weibo	14,652	1317	577.32	2104
WeChat Official Accounts	8,554	159.71	1254	22.62
Toutiao	3,642	352	1458.94	9.35
Total	26,848	18.24	17.86	-

4.3 Data Processing and Analysis

In the data processing and analysis phase, this paper used Excel to organize data. After standardizing the obtained data fields, the jieba segmentation tool was used to segment the obtained text, remove stop words, and form minimal semantic units. Subsequently, Python's IDLE development environment was used to construct the bag-of-words space for the corpus, calculate TF-IDF values, and use the K-Means algorithm for text clustering. Then, the DB index was used to test text clustering results, and finally, the optimal topic number for each day from November 22, 2017, to December 11, 2017, for "Classic Deadbeat" was obtained, and the daily topic derivative index was derived.

4.3.1 Temporal Distribution of Public Opinion Information Volume

The temporal evolution distribution of the "Classic Deadbeat" public opinion incident is shown in [Figure 2: see original paper]. Within the time range captured in this paper, there is no obvious incubation period, but in fact, the incident had already begun to ferment between April and June 2017. On June 8, 2017, the Tangshan Fengrun District People's Court ruled that defendant Huang Shufen bore primary responsibility for the accident and ordered compensation of 850,000 yuan. Therefore, the period from June 8, 2017, to November 22,

2017, can be regarded as the incubation period of this public opinion incident, during which the volume of network public opinion information was relatively small but lasted for a long time. On November 22, 2017, because Huang Shufen had been claiming she had no money and refused compensation while also refusing to communicate with Zhao Yong, Zhao Yong posted a Weibo under the username "Serious Mr. Zhao" titled "See What a Classic Deadbeat Is!" and exposed his conversation urging Huang Shufen to fulfill the legal judgment, triggering the hot spot of public opinion evolution. Public opinion quickly entered the outbreak phase (November 22-23, 2017). From November 23, 2017, public opinion evolution entered the spread phase (November 23, 2017 - December 2, 2017). The number of original information, forwards, and comments posted by users on the three major self-media platforms rose to a peak, reaching a maximum of 10,045 pieces per day. The spread phase is the main concentration stage of network public opinion incident data. By December 1, 2017, Zhao Yong's father Zhao Xiangbin died after unsuccessful rescue efforts, and public opinion entered the dissipation phase (December 2-11, 2017). Original public opinion information, comments, and forwards all decreased significantly.

4.3.2 Temporal Distribution of Optimal Topic Number The temporal evolution distribution of the optimal topic number for the "Classic Deadbeat" public opinion incident is shown in [Figure 3: see original paper]. According to the DB index of text clustering results, when the public opinion incident enters the outbreak phase, public opinion information is only divided into three different types of topics. When public opinion enters the spread phase, it begins to ferment gradually. Netizens' discussions on the incident gradually deepen and differentiate, public opinion information gradually diverges, and the number of topics reaches a peak of eight types on November 26. In the latter half of the spread phase, as the volume of public opinion information continues to decrease, the number of topics begins to decline gradually. Finally, on the second day after entering the dissipation phase (December 3), it drops to four and stops changing. At this point, it can be considered that netizens' discussion directions on the incident have completed derivation and no longer change. The number of topics and corresponding keywords for text clustering at three different time points in different stages of public opinion evolution are shown in :

Table 2 Optimal Topic Numbers and Corresponding Keywords at Different Stages

Stage	Optimal Topic Number (pieces)	Topic Keywords	Topic Description
Outbreak Phase	3	Huang Shufen, Liu Mingyue, mother-daughter, daughter, scoundrel, deadbeat, conscience, scum, buying house, anger…	Expressing anger at Huang Shufen and her daughter's behavior
		Zhao Yong, car accident, medical expenses, hit-and-run, victim, compensation, medical fees, lawsuit, driver, justice…	Sympathy and support for Zhao Yong
		Court, execution, freeze, compulsory, law, sanction, compulsory, execution, delay, intensity…	Questioning law enforcement inaction
Spread Phase	8	Huang Shufen, Liu Mingyue, mother-daughter, evil people, scoundrel, deadbeat, conscience, scum, buying house, deadbeat behavior…	Expressing anger at Huang Shufen and her daughter's behavior

Stage	Optimal Topic Number (pieces)	Topic Keywords	Topic Description
		Zhao Yong, car accident, medical expenses, hit-and-run, victim, compensation, medical fees, kindness, driver, lawsuit…	Sympathy and support for Zhao Yong
		Law, insurance, sanction, human flesh search, detention, execution by shooting, arrest, Tangshan, settlement, shame…	Offering advice for Zhao Yong
		Court, execution, freeze, justice, compulsory, violence, law enforcement, judgment, intensity, responsibility…	Questioning law enforcement inaction
		Elderly, father, Zhao Xiangbin, victim, well-being, rescue, noble, heartache, blessing, recovery…	Concern for victim Zhao Xiangbin

Stage	Optimal Topic Number (pieces)	Topic Keywords	Topic Description
		Report, network, Weibo, public opinion, law, blacklist, media, reflection, judiciary, morality…	Reflecting on effective ways to handle "deadbeat" incidents
		Employees, agents, dismissal, company, salary, executives, leaders, firing, property, resolution…	Making demands on Huang Shufen's Ping An Insurance Company
		Ping An Insurance, regret, not buying, image, contempt, senior management, products, boycott, attitude, enterprise…	Expressing anger at Ping An Insurance Company itself
Dissipation Phase	4	Huang Shufen, Liu Mingyue, mother-daughter, deadbeat, anger, scum, humanity, deadbeat behavior, conscience, despicable…	Expressing anger at Huang Shufen and her daughter's behavior

Stage	Optimal Topic Number (pieces)	Topic Keywords	Topic Description
		Zhao Yong, car accident, medical expenses, hit-and-run, victim, compensation, medical fees, driver, elderly, Zhao Xiangbin…	Sympathy and support for Zhao Yong
		Weibo, network, law, insurance, sanction, public opinion, media, morality, report, judiciary…	Offering advice for Zhao Yong
		Court, execution, freeze, justice, compulsory, delay, law enforcement, judgment, intensity, responsibility…	Questioning law enforcement inaction

Through analysis of topic keywords from text clustering results, the three most critical topics throughout the "Classic Deadbeat" incident are: condemning Huang Shufen and her daughter, sympathizing with victim Zhao Yong, and questioning law enforcement inaction. After entering the spread phase, topics derived to as many as eight sub-topics. In addition to the above three topics, there are five other sub-topics: offering advice for Zhao Yong, concern for the victim, reflecting on effective ways to handle "deadbeat" incidents, making demands on Huang Shufen's Ping An Insurance Company, and questioning Ping An Insurance Company itself. Finally, after public opinion entered the dissipation phase, public opinion heat gradually decreased, and the original eight sub-topics gradually dissipated or merged into the final four topics due to content convergence, at which point topic derivation stopped.

4.3.3 Temporal Distribution of Topic Derivative Index The topic derivative index derived from the optimal topic number at different time points is shown in [Figure 4: see original paper]. According to [Figure 4: see original paper], the topic derivative index reaches its highest value during the outbreak phase when public opinion evolves rapidly and topics differentiate sharply. After entering the spread phase, the speed of new topic generation decreases daily, and the topic derivative index gradually declines, turning negative after reaching the peak topic number. On November 27, the public opinion derivative index dropped sharply from 1 to -2. This stepwise decline indicates that public opinion sub-topics shifted from growth to decrease, and public opinion content itself shifted from divergence to convergence. After entering the public opinion dissipation phase, the optimal topic number stabilized and no longer changed, and the topic derivative index remained stable at a negative value with its absolute value continuously decreasing. Combining [Figure 3: see original paper], , and [Figure 4: see original paper], it is not difficult to find that in the "Classic Deadbeat" public opinion incident, public opinion topics, like the volume of public opinion information itself, experienced a process of outbreak—development—convergence—stability during the outbreak-spread-dissipation phases.

4.4 Discussion of Experimental Results

In previous research combining clustering algorithms and public opinion analysis, the effectiveness and scientific nature of the algorithms themselves were mainly calculated through the accuracy of clustering algorithms. The calculation formula for the accuracy (Precision) evaluation standard is as follows:

Precision = Ncorrect / Ntotal

where Ncorrect refers to the number of documents correctly clustered by the clustering algorithm that are consistent with actual categories, and Ntotal is the actual number of documents clustered.

The empirical analysis data does not contain an objective definition method for sub-topic categories and quantities. To verify the clustering effect obtained according to the DB index of text clustering validity proposed in this paper, the author additionally collected 2,240 Weibo comment messages from four categories—international politics, food, education, and medical care—in May 2019. In actual calculation processes, to verify the accuracy of the clustering algorithm multiple times, Weibo data from international politics, food, and education were first used as Dataset 1, and then all four categories of Weibo comment data were used as Dataset 2. For Dataset 1 and Dataset 2, the DBSCAN algorithm, hierarchical clustering algorithm, and the improved K-Means algorithm with DB index introduced in this paper were compared in terms of accuracy. The experimental results are shown in :

Table 3 Comparison of Text Clustering Algorithm Accuracy

Dataset	DBSCAN	Hierarchical Clustering	Improved K-Means	Average Accuracy
Dataset 1	65.3%	69.2%	75.7%	70.1%
Dataset 2	53.9%	63.6%	67.3%	61.6%

This part of the experimental results shows that compared with traditional DBSCAN and hierarchical clustering algorithms, the improved K-Means algorithm introducing the clustering validity DB index can more accurately classify public opinion information. Therefore, the optimal sub-topic number and clustering results at each moment in the empirical analysis are relatively objective and can be regarded as correct clustering of sub-topics at that moment. The public opinion derivative index derived from this clustering result thus has good validity and can be considered objective in describing the phenomenon of public opinion derivation.

5. Conclusion and Future Work

This paper theoretically distinguishes the concepts of public opinion evolution and public opinion derivation, and explores the specific meaning of public opinion derivation using text clustering and text clustering validity indexes. At the practical level, using the "Classic Deadbeat" incident as a data source, by discussing public opinion evolution events, public opinion heat, text clustering results at different stages of public opinion evolution, and the public opinion derivative index, as well as comparing the accuracy of traditional DBSCAN algorithm, hierarchical clustering algorithm, and the improved K-Means algorithm introduced in this paper, the feasibility of the public opinion derivative index in describing the phenomenon of public opinion derivation is confirmed. Meanwhile, according to the changes in the public opinion derivative index in the empirical analysis, it is known that: during the outbreak phase, the topic derivative rate is highest and then gradually declines. This phase witnesses the most intense public opinion derivation, with sub-topics showing explosive growth. The public opinion derivative index shows a stepwise decline during the public opinion spread phase, after which it remains negative. Sub-topics of public opinion begin to decrease gradually, and public opinion content itself shifts from divergence to convergence. After entering the dissipation phase, the number of sub-topics stabilizes, the public opinion derivative index remains negative and continuously approaches zero. As a measure of public opinion derivation rate and a discriminant standard for public opinion derivation, the public opinion derivative index provides a completely new perspective for public opinion supervision and early warning.

In the next stage of research, the author will apply the public opinion derivative index to conduct cross-comparisons of public opinion incidents in different fields and types to determine whether public opinion topics in different public opinion incidents all follow the development pattern of outbreak—development—convergence—stability. Another research direction is to combine multimedia

recognition, convert multimedia public opinion into text, conduct text clustering together with text public opinion, and compare whether there are differences with pure text public opinion.

References

[1] Deng Ruoyi. On the Change of Self-Media Communication and Public Sphere[J]. Modern Communication (Journal of Communication University of China), 2011(4): 167-168. [2] Gao Bin, Wang Lancheng. Research on Dissemination Model and Analysis Method of Network Derivative Public Opinion[J]. Information Theory and Practice, 2019, 42(3): 166-170, 165. [3] SAITO K, OHARA K, YAMAGISHI Y, et al. Learning diffusion probability based on node attributes in social networks[C]//International symposium on methodologies for intelligent systems. Warsaw: Springer, 2011: 153-162. [4] WATTS D J, DODDS P S. Influentials, networks, and public opinion formation[J]. Journal of consumer research, 2007, 34(4): 441-458. [5] Lan Yuexin, Dong Xilin, Zeng Runxi, et al. Research on Network Public Opinion Derivative Effect Model from the Perspective of Information Alienation[J]. Journal of Intelligence, 2015, 34(1): 139-143, 149. [6] Chen Fuji, Chen Ting. Research on Network Public Opinion Derivative Effect Based on SEIRS Dissemination Model[J]. Journal of Intelligence, 2014, 33(2): 108-113, 160. [7] Yin Xicheng, Zhu Hengmin, Ma Jing, et al. Coupled Network Model of Microblog Public Opinion Topic Dissemination—Analyzing Topic Derivative Characteristics and User Reading Psychology[J]. Information Theory and Practice, 2015, 38(11): 82-86. [8] Wang Lijun, Dai Jianhua. Research on Quantitative Analysis Method of Network Public Opinion Derivative Chain[J]. Information Science, 2016, 34(7): 59-63. [9] Jin Xiaohong, Wang Qiang, Fu Hong, et al. Construction and Empirical Research of Thematic Event Public Opinion Index—Taking Food Safety Theme as an Example[J]. Information Theory and Practice, 2016, 39(12): 103-108. [10] He Enfeng, Zhuang Linyuan, Xu Wengen. Construction and Application of Network Public Opinion Potential Influence Indicator System[J]. Journal of Intelligence, 2014, 33(1): 114-119. [11] Deng Shangmin, Dong Yaqian. Research on Construction of University Network Public Opinion Safety Assessment Indicator System Based on AHP[J]. Journal of Intelligence, 2012, 31(8): 31-36. [12] Chen Fuji, Huang Jiangling. Literature Review on the Evolution of Network Public Opinion in China[J]. Journal of Intelligence, 2013, 32(7): 54-58, 92. [13] Zhou Xin, Huang Wei, Teng Guangqing, et al. Analysis and Reconstruction of Network Public Opinion Dissemination Model[J]. Information Theory and Practice, 2016, 39(12): 25-30. [14] Huang Wei, Wang Jiejing, Zhao Jiangyuan. Research on Measurement of Weibo Public Opinion Information Aging[J]. Information and Documentation Services, 2017(6): 6-11. [15] Zhou Kaile, Yang Shanlin, Ding Shuai, et al. Review of Clustering Validity Research[J]. Systems Engineering—Theory & Practice, 2014, 34(9): 2417-2431.

Author Contributions

Huang Wei: Paper writing guidance;
Zhu Zhenyuan: Topic selection and paper writing;
Xu Yejing: Data collection and organization;
Sun Yue: Paper proofreading.

Establishment of Public Opinion Derivative Index: An Empirical Study in China

Huang Wei, Zhu Zhenyuan, Xu Yejing, Sun Yue
School of Management, Jilin University, Changchun 130022

Abstract: [Purpose/significance] During the evolution of public opinion, the derivation of public opinion could possess significant value for the forecasting and warning of public opinion both theoretically and empirically. [Method/process] To investigate the mechanism of public opinion derivation, this paper conducted the study using text clustering and DB cluster validity index. It proposed certain standards to judge the occurrence of public opinion derivation and its according velocity index. Furthermore, this paper used a well-known public opinion incident called "Classic Deadbeat" to conduct empirical research. [Result/conclusion] The result of empirical study shows that: the derivative index reaches its climax during emergence phase and declined thereafter. The number of sub-topics reaches its climax during integration phase and then declined thereafter; when the number of sub-topics decreased, the derivative index become negative, indicating that the public opinion becomes stabilized. When the public opinion incident reaches the disappearance phase, the number of subtopics becomes stable and the derivative index remain negative but approach zero. The study of derivative index of public opinion offers a new angle to study public opinion observation and prediction.

Keywords: public opinion derivative; derivative index; text clustering; cluster validity indexes

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