Enhanced Latent Semantic Analysis and Support Vector Machine Algorithms for Public Opinion Early Warning of Emergency Security Incidents: Postprint

Authors: Tian Shihai, Lü Deli

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】To address the limitations of current early warning systems that primarily focus on enterprises and regulatory authorities while neglecting online public opinion—resulting in weak effectiveness, lack of transparency and sensitivity, and frequent sudden safety issues that cannot be timely resolved—this paper proposes a novel public opinion early warning model.

【Method】By utilizing meta-search technology to extract public opinion information, adding baseline offset values to optimize the tendency weights of sentiment features, and incorporating correction factors to improve the Latent Semantic Analysis and Support Vector Machine (LSA+SVM) algorithm, a public opinion classification and early warning model is constructed.

【Results】Using multiple sudden safety incidents as case studies, simulation experiments are conducted using Matlab. The results demonstrate that the proposed public opinion early warning model is feasible and responsive, achieving an accuracy of 85.75% when the semantic dimension is 10.

【Limitations】This approach is more effective for safety incidents that generate public attention and discussion.

【Conclusion】The improved algorithm is suitable for public opinion early warning and can offer reasonable recommendations for enterprises and regulatory authorities to take effective early warning measures in a timely manner based on classification results.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 2, 2017

Improved Latent Semantic Analysis and Support Vector Machine Algorithm for Early Warning of Public Opinion on Sudden Safety Incidents

Tian Shihai, Lyu Deli

(School of Management, Harbin University of Science and Technology, Harbin 150040, China)

Abstract

[**Objective**] Existing early warning systems primarily focus on enterprises and regulatory agencies while neglecting online public opinion, resulting in weak early warning capabilities, lack of transparency and sensitivity, and frequent occurrences of sudden safety problems that cannot be addressed promptly. To address this gap, this study proposes a novel public opinion early warning model. [**Methods**] The model employs meta-search technology to mine public opinion information, optimizes the weight of sentiment feature terms by adding a baseline offset value, and introduces a correction factor to improve the Latent Semantic Analysis and Support Vector Machine (LSA+SVM) algorithm, thereby constructing a public opinion classification and early warning model. [**Results**] Using multiple sudden safety incidents as case studies, Matlab simulation experiments demonstrate the feasibility and rapid response of the proposed model, achieving an accuracy rate of 85.75% when the semantic dimension is set to 10. [**Limitations**] This method proves more effective for safety incidents that attract significant public attention and discussion. [**Conclusions**] The improved algorithm is suitable for public opinion early warning and can provide reasonable recommendations for enterprises and regulatory agencies to take timely and effective measures based on classification results.

Keywords: Latent Semantic Analysis; Support Vector Machine; Public Opinion Early Warning; Sentiment Orientation Analysis

Classification Number: G203

Introduction

Online public opinion is characterized by rapid dissemination, multiple channels, and wide coverage. Its propagation, diffusion, and fermentation regarding hot-button and sudden incidents play a crucial role in corporate decision-making and management. However, public opinion information is often chaotic, carries strong emotional coloring and noise, and may even threaten corporate survival and development. Therefore, how to properly utilize online public opinion and conduct timely classification and early warning of relevant information should be a key focus for both enterprises and scholars.

Extensive research has been conducted on public opinion early warning both domestically and internationally. Wu et al. [?] constructed a network group behavior model through Agent modeling. Li et al. [?] applied artificial neural networks to predict product safety. Wang Lancheng [?] analyzed the functions of public opinion intelligence and designed an intelligence support system architecture for emergency response. Papetti et al. [?] proposed a multi-factor, multi-data source public opinion early warning model, validating through multiple cases that the new model could ensure early warning accuracy while reducing warning time and source data requirements. Dong Kaixin et al. [?] analyzed role indicators and identified opinion leaders through subgroup mining to propose recommendations for public opinion mechanisms. Chen Fuji et al. [?] effectively predicted public opinion event development trends by establishing opinion interaction mechanisms.

In summary, most existing public opinion early warning research models and predicts early warning measures at the macro level. However, the distribution of sentiment feature words is unbalanced across the entire model, and semantic dimensions are complex. Therefore, more precise classification methods must be investigated from the perspective of optimizing semantic dimensions and speed. This paper establishes a public opinion classification and early warning model by improving Latent Semantic Analysis (LSA) and Support Vector Machine (SVM) algorithms to enhance the accuracy of orientation prediction, improve classification efficiency, and strengthen public opinion situation awareness. This ensures enterprises can take proactive and effective measures before risks escalate further while addressing their own problems, innovating products, and adapting to market requirements based on public opinion feedback.

Sudden safety incidents have four distinctive characteristics: (1) They are significantly influenced by national policies and regulations, which are updated in real-time across control, inspection, and management aspects as safety incidents occur and technology advances—a critical factor for enterprise early warning. (2) They are highly sudden yet slow-burning, with long fermentation periods. Safety incidents often originate from unexpected events, spread rapidly, and attract substantial attention. They typically involve corporate production management systems and industry inspection mechanisms, leading to prolonged periods of dissemination, fermentation, and 沉淀. (3) They attract broad public attention and have strong coverage. Since online content is closely related to daily life and individual safety, the public tends to devote more attention until incidents are resolved. (4) They have severe impact and can be devastating to enterprises, as seen in the "Bawang Carcinogen" controversy and the "Melamine" incident. Consequently, enterprises should allocate more resources to crisis early warning. Based on these characteristics, public opinion classification and early warning models must minimize the semantic dimension of public opinion orientation to enable rapid capture of public sentiment during the initial stages of safety incidents while maintaining excellent combination and classification capabilities to timely rate incidents and accurately determine their orientation.

LSA can eliminate deviations caused by synonyms and polysemous words in text analysis, obtain more accurate text vectors, and simplify text vectors to improve computational efficiency. SVM, as a classifier with excellent generalization ability, is widely applied and can be extended to other machine learning problems such as function fitting [**?**]. Therefore, this paper selects the LSA and SVM algorithm combination to meet classification and early warning requirements, with appropriate improvements to better align with the characteristics of public opinion subjects. The construction of the public opinion classification and early warning model mainly includes the following steps: first, analyzing the public opinion classification and early warning process; second, determining and correcting sentiment feature word weights and improving the LSA+SVM algorithm; and finally, implementing the algorithm model.

2. Construction of Public Opinion Classification and Early Warning Model

2.1 Public Opinion Classification and Early Warning Process

The public opinion classification and early warning process consists of three main stages: information capture, orientation determination, and public opinion classification. Information capture utilizes meta-search technology and Nutch crawlers to perform simple noise reduction, cleaning, and segmentation on captured data to extract sentiment feature words. This study focuses on orientation determination and classification. The public opinion classification and early warning process is illustrated in Figure 1 [Figure 1: see original paper].

2.2 Determination and Improvement of Sentiment Feature Word Weights for Public Opinion Classification

Hot keywords related to safety incidents are selected, and the open-source tool Nutch crawler is used to mine public opinion corpora, obtaining a relevant URL list of public opinion. Combining standards from the HowNet sentiment analysis lexicon, sentiment orientation feature words are extracted, primarily adjectives, adverbs, and nouns. Sentiment feature term texts are vectorized and stored in the following format: ia represents the sentiment feature item for public opinion classification; T represents the time when the feature item was obtained; t represents the publication time of the relevant text where the feature item was obtained; r is a binary field indicating whether the URL was reposted; and W represents the importance weight value of the source webpage for the feature item. When r is "yes," W takes the weight value of the sentiment feature item.

Considering that the degree of information source influence and the semantic orientation consequences of information will significantly affect the weight of relevant public opinion feature items, let the importance degree of sentiment feature items appearing in this public opinion text vector be ik_{tfidf} [?]. The term frequency ik_{tf} represents the frequency of sentiment feature item ia. ik_n represents the number of occurrences of feature item ia, and ik_{tf} needs to be calculated in conjunction with the total number of feature items iN appearing in the entire text vector. The inverse document frequency idf represents the inverse document frequency of sentiment feature item ia, i.e., words that appear less frequently in the entire article but have distinct features. Therefore, the inverse of the number of feature items ia is calculated, and the logarithm of this value is used:

However, in real public opinion texts, long sentences involve many adverbs and nouns with more pronounced sentiment orientation, causing the weight values of feature items to favor long texts. This leads to the log function becoming zero, losing its judgment capability. Meanwhile, safety incidents commonly use empirical coefficients containing semantics and modality to highlight important feature items, and temporary releases or improvements of national policies and regulations significantly impact relevant industries. Therefore, to solve the problem of the log function becoming zero, its value is increased by 0.01. To address temporary policy and regulation impacts, a baseline offset value $offset$ [?] is added to improve the main weight value, yielding the public opinion classification sentiment feature item weight calculation formula:

Through the public opinion classification sentiment feature item weight formula, the weight values of vectorized texts are calculated and stored to facilitate the next step of vector space operations.

2.3 Vectorization and Classification of Sentiment Feature Words for Public Opinion Classification

Sentiment feature words for public opinion classification are stored as single text vectors that do not belong to the same concept space, resulting in excessively high space dimensions that require dimensionality reduction for combination and classification. The basic process of the sentiment feature word classification method based on the improved LSA+SVM algorithm is shown in Figure 2 [Figure 2: see original paper].

Text preprocessing involves weight value calculation and improvement. LSA segments public opinion classification texts into different local feature spaces through singular value decomposition, avoiding interference from noise terms such as polysemous words and synonyms, making the meanings expressed by public opinion sentiment feature words more explicit and perceptible. The vectorized public opinion feature word space vector is decomposed and stored in an $m \times n$ matrix format:

For preliminary processing of the sentiment feature item matrix, if multiple public opinion sentiment feature words are synonyms with high semantic relevance, they are classified into the same category. Consequently, the probability of different category feature words being synonymous is lower. Thus, matrix A is decomposed into a combination form of multiple matrices of different categories, as follows:

where U and V are the left and right singular vector matrices of A^{TA} and AA^T, respectively, and S is the singular value matrix of matrix A. Then $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$. Singular Value Decomposition (SVD) is used to compress the entire USV^T space, obtaining a k-rank matrix in the following form:

The specific decomposition process is shown in Figure 3 [Figure 3: see original paper].

The similarity relationship of feature words is calculated through the inner product between row vectors of matrix A_k:

The TSS calculated from $U_k S_k S_k^T U_k^T$ represents the inner product relationship between rows i and j, reflecting the similarities and differences between two vectors, where k represents the dimension after dimensionality reduction. The new text vector is obtained and sent to the SVM classification module for correlation-based classification.

2.4 Improvement of LSA+SVM Algorithm for Public Opinion Classification

Sudden safety incidents have long fermentation periods, broad audiences, and significant impacts on corporate reputation. Ordinary classifiers struggle to determine their sentiment orientation and danger level, requiring the addition of a correction factor iaO [**?**] in their feature word local matrices. The correction factor primarily uses the simultaneous appearance of sentiment words fa and degree adverbs ga in the local feature vector as a benchmark, multiplying their weights to obtain a value that serves as the priority judgment standard for matrix severity. The calculation method is:

In $A_k = U_k S_k V_k^T$, S_k is rearranged using the correction factor iaO and the local matrix obtained after singular value decomposition. Simultaneously, based on the weight values of safety feature words, a baseline offset value is added to deviate from the original arrangement and trend. Several linear relationships of singular values are simulated as a regression jump curve with correlation, as shown in Figure 4 [Figure 4: see original paper].

When $iaO > 0$, the correlation jump curve is shown in Figure 4(a), indicating that the local matrix of this public opinion classification has strong positive meaning, generally appearing in comments or texts that are optimistic about the event or even helpful to corporate brand image. When $iaO = 0$, the curve shows no obvious jump, generally fluctuating near the horizontal axis, and the orientation of this local matrix tends to be neutral. Such comments are more narrative, containing no obvious criticism or support. When $iaO < 0$, the correlation jump curve is shown in Figure 4(b), indicating that this local matrix has strong negative meaning, suggesting that the text has strong reactions to the event and obvious criticism of the enterprise. The closer the iaO value is to 1 or -1, the more severe the sentiment orientation of the feature word.

2.5 Implementation Process of Improved LSA+SVM Algorithm

(1) Training Algorithm Implementation

A large number of public opinion classification training texts are selected to train the improved LSA+SVM algorithm, forming a standard public opinion classification and early warning parameter model to obtain three basic parameters: χ, α, and β, representing the penalty function coefficient, linear maximum margin, and kernel function coefficient, respectively. The specific training algorithm process [?] is as follows:

Input: Feature word vector set $\{ia_1, ia_2, \dots, ia_n\}$, baseline offset value $offset$
Output: Classification parameter model $\{\chi, \alpha, \beta\}$

(2) Testing Algorithm Implementation

The testing algorithm combines the parameter model and SVM classifier to classify the sentiment orientation of new feature item texts. It first divides them into two levels based on the positive or negative nature of the correction factor, then relies on weight values to divide them into five levels: Special-level public opinion (S-level), Heavy-level public opinion (A-level), Medium-level public opinion (B-level), Light-level public opinion (C-level), and Needs-attention public opinion (D-level). Positive public opinion is incorporated into corporate feedback and innovation information as (P-level) [?]. The testing algorithm process is as follows:

Input: Feature word set to be classified $\{ia_1, ia_2, \dots, ia_n\}$, baseline offset value $offset$
Output: Classification result $\{S, A, B, C, D, P\}$

Through correction based on baseline offset values and correction factors, along with extensive text training, the model can more accurately and efficiently determine the crisis situation of real-time safety incidents and provide timely feedback to enterprises, achieving the early warning purpose.

3. Implementation and Simulation of Public Opinion Classification and Early Warning

To ensure more accurate public opinion classification, this section discusses the practical application of the LSA+SVM algorithm using three sudden safety incidents from different domains: the Mengniu aflatoxin incident (Incident 1) representing food safety, the Baidu "Putian Department" incident (Incident 2) representing internet user security, and the Tianjin Binhai chemical plant leakage incident (Incident 3) representing production safety. First, according to public opinion categories, a "hot words + public opinion words" format is designed. Using Python and meta-search technology, crawlers are deployed across various search engine hot news sections to mine these fields, obtaining over 900 relevant articles and comment URL lists for the series of incidents, as shown in Figure 5 [Figure 5: see original paper].

After simple deduplication and noise reduction of the crawled documents, the Institute of Computing Technology, Chinese Academy of Sciences' ICTCLAS segmentation system and the LSA+SVM algorithm are selected to obtain the basic classification parameter model through cross-validation. The algorithm implementation is based on the Windows 7 operating system, with Matlab 2012b as the simulation software. The training selects the Radial Basis Function (RBF) kernel function and uses cross-validation to determine the optimal parameter and classification models.

The trained kernel function coefficient is approximately 0.431, and the penalty function coefficient is 0.424462. A total of 324 effective feature vectors with negative orientation and 198 effective feature vectors with positive or neutral orientation are obtained. According to the movement of the final classification jump curve, it is found that the jump curve tends to be more negatively correlated, indicating negative public opinion orientation. Comparative experiments are conducted on the classification model under different semantic dimensions, using the accuracy of document orientation under different parameters as the basic performance metric:

where P represents the total number of positive documents selected, PP represents the number of documents that were positive at selection and remain positive after classification when $iaO > 0$; similarly, NN represents the number of documents that remain negative after classification. Three groups of different semantic dimension values are randomly selected, with k values of 5, 10, and 15 for accuracy calculation. The results are shown in Figure 6 [Figure 6: see original paper].

The results show that performance is best when $k = 10$, with accuracy reaching 87.25%, which can efficiently reflect the relevant characteristics of texts. Too low a dimension easily leads to result bias, while too high a dimension easily causes semantic confusion and inaccurate grading [**?**]. The classification algorithm implementation results are shown in Figure 7 [Figure 7: see original paper].

The final effective documents and comments are classified according to the correction factor value. Positive public opinion can all be classified as P-level, while negative public opinion requiring alerts—levels S, A, B, C, and D—are equally divided in the $(-1, 0)$ interval. Thus, $S \in [-1, -0.8]$, $A \in [-0.8, -0.6]$, $B \in [-0.6, -0.4]$, $C \in [-0.4, -0.2]$, $D \in [-0.2, 0)$. However, considering that the $(0, 0.1)$ interval, although belonging to positive public opinion, shows unclear orientation, this interval is classified into D (needs attention) [**?**]. Excluding duplicate and training documents, the positive and negative document ratios for the three sudden safety incidents are shown in Table 1 .

Based on the positive and negative document ratios, the public opinion orientation of different sudden safety incidents can be preliminarily observed. Here, negative and neutral documents are selected to further summarize the number of public opinion documents at different levels, to determine the urgency of each sudden safety incident and what measures enterprises should take. The level classification and main public opinion vocabulary for the three sudden safety incidents are shown in Table 2 .

Table 2 reveals that the public opinion levels of these three sudden safety incidents all belong to S-level (Special-level) public opinion, with the Binhai chemical plant explosion and leakage incident being the most severe. All require high attention from enterprises or regulatory agencies and prompt response.

4. Conclusions and Recommendations

To address the current issues of narrow early warning scope, weak transparency, and untimely response for sudden safety incidents, this study incorporates external entities into network public opinion early warning research. Based on real-time mining of hot events and keywords, combined with relevant characteristics of public opinion classification, baseline offset values for weights are added to improve the LSA+SVM algorithm. Through the positive and negative values of correction factors, public opinion orientation determination and early warning classification are conducted:

(1) When the correction factor is negative and ranges in $[-1, -0.4)$, it is determined as S, A, and B levels based on weight ranking, representing that the public opinion source has obvious negative orientation and significant impact, requiring timely enterprise intervention.

(2) When the correction factor ranges in $[-0.4, 0.1)$, the public opinion source is determined to belong to C and D levels, representing neutral public opinion that requires monitoring and observation.

(3) When the correction factor ranges in $[0.1, 1]$, it is determined as P-level public opinion, belonging to positive public opinion that helps maintain a positive corporate image.

Using LibSVM and Matlab for simulation and accuracy calculation, the algorithm' s effectiveness is verified. The results can reflect the orientation of public opinion texts and provide accurate alert information for enterprises. Finally, the following recommendations are proposed for measures enterprises should take:

S-level (Special-level public opinion): Attach great importance and respond promptly. Enterprises need to immediately dispatch professional public relations teams, quickly identify the source of public opinion, conduct timely product recalls and compensation processing, minimize the harm of negative impacts to corporate image, and establish a responsible corporate image.

A-level (Heavy-level public opinion): Take measures to resolve the crisis. Heavy-level public opinion requires timely enterprise intervention to prevent further diffusion that could transform it into special-level public opinion. At this stage, enterprises can measure their own resources and crisis handling capabilities, integrate resources without harming current corporate interests, and avoid crisis escalation.

B-level (Medium-level public opinion): Suppress further diffusion of public opinion information. Closely monitor the status of crisis information and the direction of public opinion, appropriately guide public discourse, and activate contingency plans to ensure the event develops in a favorable direction.

C-level (Light-level public opinion): Eliminate interference information and respond actively. Provide corresponding improvement suggestions to relevant enterprise departments, continue monitoring, and respond promptly to any changes.

D-level (Needs attention): Conduct daily monitoring effectively. Conduct preliminary judgment of public opinion categories, include positive public opinion in the corporate innovation knowledge base, and record negative public opinion as a potential problem source for prevention and preparedness.

P-level (Information feedback): Serve as feedback suggestions. Since most P-level information lacks excessive emotional coloring or is primarily positive, enterprises can refer to this feedback to innovate products, strengthen management, and upgrade services, providing new ideas and opportunities for corporate development.

References

[1] Wu Peng, Yang Shuang, Zhang Jingjing, et al. Agent-Based Modeling and Simulation of Evolution of Netizen Crowd Behavior in Unexpected Events Public Opinion[J]. New Technology of Library and Information Service, 2015(7/8): 65-72.

[2] Li W, Miao D, Wang W. Two Level Hierarchical Combination Method for Text Classification[J]. Expert Systems with Applications, 2011, 38(3): 2030-2039.

[3] Wang Lancheng. Research on Emergency Information Support Based on Network Public Opinion Analysis[J]. Information Studies: Theory & Application, 2015, 38(7): 72-75.

[4] Papetti P, Costa C, Antonucci F, et al. A RFID Web-based Infotracking System for the Artisanal Italian Cheese Quality Trace Ability[J]. Food Control, 2012, 27(1): 234-241.

[5] Dong Kaixin, Fu Ying, Sun Xiaofeng, et al. Research on Early Warning Mechanism of Enterprise Public Opinion Based on Social Network Analysis —Taking Food Safety Network Public Opinion as an Example[J]. E-Business Journal, 2015, 23(8): 54-55, 57.

[6] Chen Fuji, Chen Ting. Research on Public Opinion Emergencies Evolution: Based on the Perspective of Opinion Leaders Guiding Role[J]. Information and Documentation Services, 2012, 36(2): 23-28.

[7] Xuan Yungan, Zhu Qinghua. Research on Tag Semantic Retrieval in Social Tagging System Based on LSA[J]. Library and Information Service, 2011, 55(4): 11-14.

[8] Fan Yuhua, Qin Shiyin. Optimizing Decision for Scene Classification Based on Latent Semantic Analysis[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(2): 175-182.

[9] Shang Liyuan, Tan Qingmei. Emergency Classification Based on Support Vector Machine[J]. Journal of Industrial Engineering and Engineering Management, 2014, 28(1): 119-123.

[10] Zhang Jian' e. A Chinese Keywords Extraction Approach Based on TFIDF and Word Correlation[J]. Information Science, 2012, 30(10): 1542-1555.

[11] Zhang Changli. Research on Domain-Oriented Public Sentiment Analysis Technologies[D]. Changchun: Jilin University, 2011.

[12] Gao Hongyan, Wang Jianhui. A Fuzzy Control Method with Online Self-turning Correction Factor and Its Application[J]. Microcomputer Information, 2006, 22(13): 83-84.

[13] Goñi S M, Oddone S, Segura J A. Prediction of Foods Freezing and Thawing Times: Artificial Neural Networks and Genetic Algorithm Approach[J]. Journal of Food Engineering, 2011, 84(1): 164-178.

[14] Tan Guangxing, Liu Zhenhui. A Local Latent Semantic Analysis Algorithm Based on Support Vector Machine[J]. Computer Engineering and Science, 2016, 38(1): 177-182.

[15] Sengupta A S, Balaji M S, Krishnan B C. How Customers Cope with Service Failure? A Study Does Brand Reputation and Customer Satisfaction[J]. Journal of Business Research, 2015, 68(3): 655-674.

[16] Zhu Ge, Qi Jiayin. Situation Evaluation of Online Public Opinion on Enterprise Crisis Event[J]. Information Science, 2015, 33(6): 48-53.

[17] Ma Ning, Liu Yijun. Multi-Agent Modeling of Public Opinion Evolution Based on SuperNetwork Analysis[J]. Journal of Systems & Management, 2015, 24(6): 785-804.

Supporting Data

The supporting data is self-archived by the authors, E-mail: lvdelixx@126.com.

[1] Tian Shihai, Lyu Deli. pythonurl1.txt. Python crawler mining code.
[2] Tian Shihai, Lyu Deli. wo3url.csv. Mined URL list.
[3] Tian Shihai, Lyu Deli. lsasvm.csv. Segmented sentiment orientation feature

word matrix.
[4] Tian Shihai, Lyu Deli. lsasvmdepart.rdf. Feature word vector matrix calculated by improved LSA+SVM algorithm.
[5] Tian Shihai, Lyu Deli. orien.xls. Document orientation accuracy list.

Author Contributions

Tian Shihai: Proposed research ideas, designed research framework, revised final version of paper.
Lyu Deli: Implemented research process, acquired data, conducted experiments, wrote paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Received Date: 2016-08-29
Revised Date: 2016-10-25

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