Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2069562, 10 pages
https://doi.org/10.1155/2022/2069562

Research Article

Corpus-Based Data Acquisition and Topic Analysis of Chinese-Related Public Opinion in Western Media

Xiulian Zhao

Jiangsu University of Technology, Jiangsu 213100, China

Correspondence should be addressed to Xiulian Zhao; zhaoxiulian@jsut.edu.cn

Received 9 May 2022; Revised 6 June 2022; Accepted 8 June 2022; Published 5 July 2022

Academic Editor: Kalidoss Rajakani

This paper sorts out and discusses the different methods used in the analysis of China-related public opinion, builds a theoretical system of public opinion analysis methods on this basis, conducts case analysis and research on hot public opinion combined with corpus, and selects the semantic analysis method from the technical level supporting the public opinion analysis method. From the perspective of related technologies involving intelligent analysis methods, comparative analysis and improved applications are given, which provides an effective analysis basis for government public management departments to fully grasp and guide China-related public opinion. Sentiment tendency analysis is also an important direction of Chinese-related public opinion research. This paper sorts and analyzes China-related public opinion from the characteristics, applications, and text processing methods of different granularities and builds and improves the annotation model for the smallest granularity. The frequency, trend, and evolution characteristics of China-related public opinion events are sorted out and analyzed, and the trend analysis method is used to analyze the evolution trend of attention of public opinion events. The distribution of individual opinion acceptance, trust threshold, and opinion leaders are simulated by experiments. The impact on the evolution of China-related public opinion was determined. The experimental results show that the effect of the method proposed in this paper is improved by about 10% compared with the direct use of statistical learning methods for emotional orientation analysis. The SVM method also has obvious advantages over the Bayesian method, especially the combination of the SVM method and the bigram method which gives the best results.

1. Introduction

With the development of Internet technology and the popularization of network applications, the Internet has become an important source for people to obtain information and also an important channel for people to disseminate information and express their opinions. Understanding social conditions and public opinion through the Internet and paying attention to the trend of public opinion have important practical significance for promoting social harmony and stability and promoting social democracy and legal system construction. The network information is vast and mixed, and manual identification and judgment are not enough. How to use computer network technology, artificial intelligence technology, and data mining technology to effectively mine and analyze China-related public opinion information has become a new research hotspot [1]. How to identify the hot topics that the public is concerned about and classify them effectively, how to judge whether the public's attitude towards social events is positive or negative, how to analyze and grasp the volatility of hot social events, etc. are the areas of interest. The key issues to be solved have important scientific significance for understanding and guiding China-related public opinion [2]. Regarding the analysis of the evolution model of China-related public opinion, the main body of China-related public opinion is generally regarded as independent and scattered particles. The evolution of China-related public opinion is analyzed by studying the interaction between individuals in social groups. This research is mainly to systematically sort out and analyze

the theoretical methods and supporting technologies related to the analysis of China-related public opinion. Due to the comprehensiveness and intersectionality of this research direction, this paper will comprehensively use the method of combining various disciplines and follow the steps from theory to model. Research on the technical route of simulation was performed.

With the increasing growth and complexity of network information resources, network content analysis has begun to expand its analysis scope and extension. According to the interpretation of the meaning of content analysis, we can understand that content analysis is a quantitative analysis method based on qualitative analysis. Its analysis target is information content and is aimed at extracting effective information features by means of systematic and in-depth analysis of information content [3]. The information composed of words and sentences is expressed in a quantitative form to realize the quantitative processing and expression of statistical results. Its processing objects are not limited to explicit information and can also process latent or implicit information that is not easy to be detected. The most prominent difference between implicit information and explicit information is that it does not directly express the internal relationship between events but reveals its essence and attributes by relying on a large number of external characteristics. No matter what kind of information is researched, it is necessary to accurately grasp its essence. It can be said that the process of information analysis is a process of accurately and deeply grasping the nature and attributes of information. Content analysis is fundamentally a quantitative analysis method based on qualitative analysis.

China-related public opinion often contains a lot of information closely related to people's livelihood issues and social issues. Using the content analysis method to analyze it deeply and systematically can more accurately predict the distribution of public opinion and future development trends [4]. The content analysis method can express the hidden information that is not easy to be detected in the public opinion information, and this is also an important application of this method. China-related public opinion information can clearly reflect the positions and propositions of public opinion makers and information disseminators, but sometimes people with ulterior motives will maliciously create and spread public opinion, such as spreading online rumors. In this regard, the content analysis method can be used to conduct in-depth and detailed analysis of the opinions expressed by the majority of netizens on the Internet, and on the basis of trying to clarify their intentions, accurately determine their emotional attitudes and positions, and provide them to relevant institutions for further comparison, research to speculate on the true intentions of opinion makers and disseminators was conducted. In addition, for China-related rumors, we can rely on existing technologies and methods to explore the root cause and take effective measures and methods to minimize their adverse effects. Therefore, content analysis method can infer the intentions and emotional tendencies of the main body of Chinese-related public opinion information dissemination, which is the second application of content analysis method [5]. China-related public opinion is constantly changing, and it has always maintained a close interactive relationship with public opinion events and netizens. The accurate and detailed discussion and rigorous and comprehensive analysis of the public opinion information in each time period can evaluate the correlation between the characteristics of the public opinion content and the creator of the public opinion, thereby clarifying the source of the public opinion information [6].

Sections of this paper are arranged as follows: Section 2 introduces the related scholars' research on public opinion analysis and sentiment tendency, Section 3 introduces the support vector machine technology and semantic text tendency subtechnology in the analysis technology of public opinion tendency analysis, Section 4 makes various comparative experiments on the method of this paper through specific data, and Section 5 is the summary of the full text.

The innovation of this paper is as follows: it analyzes the connotation and characteristics of China-related public opinion events and sorts out and analyzes the occurrence frequency and trend of China-related public opinion events. Taking 100 web pages as the data source, using them as the semantic knowledge resources of the system, by introducing the influence factor of the relative position of the sememe and the depth influence factor for improvement, and applying them to the calculation process of sentence similarity and paragraph similarity, the calculation results of the improved method are more accurate and in line with the reality through relevant experiments.

2. Related Work

The study of public opinion is inseparable from semantic analysis, and there is a close relationship between semantic interpretation and context. Only by correctly understanding the context can we accurately interpret the semantic meaning. No matter which edition of the dictionary is, the explanations given for words are out of the specific context, showing a strong isolation. When a word is applied to a specific sentence or paragraph, it becomes an indispensable and important part of the document, and together with other words, it constitutes a complete context, which is the context we usually talk about. Context clarifies the direction for semantic interpretation, which objectively and accurately expounds the specific impact of other factors, not including language, on the choice of language description forms [7, 8].

Li et al. first introduced the method of machine learning into sentiment classification and used Bayesian classifier, maximum descendant model, and support vector machine to classify movie reviews. The experimental results show that support vector machines have achieved the best classification results [9]. Tian and Shen compared more aspects of the applied machine learning method, including the calculation method of weight and the location information of feature selection, such as words at the front, middle, or near the end of the document, which are often used as classification features; part of speech annotated information and adjectives are often considered to have a strong ability to represent emotions [10]. Xue-Feng and Chen proposed a topic sentiment mixture model for topic sentiment analysis on weblog, focusing on how to extract propensity semantic

information [11]. H. Wang and G. Wang explained that semantics is completely reflected by contextual relations, and based on this, relevant semantic analysis is carried out [12]. Cui applies semantic analysis to movie reviews and determines the recommendation index based on the results of semantic analysis and contextual structure [13]. Hu and Hong collected the corpus information on Twitter as a corpus for semantic analysis and used the SVM classifier to judge the category in the process of singular value decomposition of the matrix, which effectively optimized the effect of web text classification [14]. Wang and Sun used semantic analysis to judge the tendency of various opinions in China-related public opinion and used a mining algorithm to solve the problem of text classification [15]. Ha proposed a method combining semantic analysis and geographic location unsupervised judgment of emotional tendencies to automatically assess the engagement of social network users in different countries in events with global influence [16]. Jeong et al. used optimal segmentation theory and Moran's I index to construct an evolution model of unconventional emergencies involving China based on the continuous and clustering characteristics of search engine attention data and conducted empirical research to reveal the law of evolution of China-related public opinion in unconventional emergencies [17]. Fan draws on the research method of system dynamics to analyze the evolution mechanism of China-related public opinion in "NIMBY" conflict events and builds an evolution model of China-related public opinion in "NIMBY" conflict events, which provides an important tool for controlling the spread of China-related public opinion and resolving conflict events [18]. Liu et al. proposed an evolution analysis method of network public opinion based on the guiding role of opinion leaders, which can simulate the evolution process of China-related public opinion and can reasonably predict the trend of network public opinion [19]. In the research of sentiment classification using machine learning methods, Chen and Huang selected words with semantic tendency as features and correctly processed negative words and used binary values as feature weights to improve the classification accuracy. Luo proposed to use frame technology to extract sensitive elements from different aspects of the description of the report to form a set of sensitive elements. As a classification system, find out the key sentences containing these elements in the report and according to the information structure and position provided by the sentence. Propensity calculation is carried out by concept library, etc. [20]. Jia combines the sentiment analysis based on lexical level and sentence level. In the lexical-level text sentiment analysis method, a conditional random field is used to identify sentiment words and judge their polarity. At the sentence level, a text orientation discrimination model based on maximum entropy is used [21].

There have been many discussions on the tendency analysis of public opinion information using the above methods. These explorations of Chinese public opinion are still carried out around two aspects: first is to extract effective sentiments from sentiment analysis texts, especially Chinese public opinion, and second is to continuously improve the performance of the propensity classifier.

3. Related Technologies of Public Opinion Tendency Analysis

3.1. Analysis of China-Related Public Opinion Tendency Based on Support Vector Machine. Due to the diversity and complexity of the expressions of China-related public opinion and the diversity and complexity of the semantic grammar for describing China-related public opinion texts, it is necessary to establish a general mechanism to enable accurate analysis methods of different China-related public opinions. The combination of public opinion semantics and grammar can complete the task of accurate analysis of Chinese-related public opinion. Public opinion refers to the social and political attitudes to social managers and their political orientations that are generated and held by the occurrence and changes of intermediary social events in a certain historical stage and social space. Simply put, it is the social and political attitude of the people. It is the sum of the attitudes, opinions, and emotions expressed by the majority of the masses towards various phenomena, problems, and events in the society.

The early work of this method generally regards vocabulary as the basic processing unit, first calculates or analyzes the tendency of praise and disapproval of the vocabulary, and then adds the semantics of the vocabulary to calculate the tendency of the text. There are two methods for studying the tendency of Chinese-related public opinion based on semantics. The first method is to first select phrases that can reflect the subjective color part of speech, such as adjectives, for text analysis, and then judge the tendency of each extracted phrase, assign a tendency value to each, and finally find the corresponding value of all phrases. The sum of the propensity values can be used to judge the overall propensity of China-related public opinion [22]. The second method is to establish a preference semantic pattern library in advance, and the pattern library comes with a commonly used preference dictionary. Then, pattern matching is performed between the text to be analyzed and the semantic pattern library, and finally the sum of the corresponding tendency values is obtained, so as to obtain the tendency of the entire Chinese-related public opinion [23].

Support vector machine, referred to as SVM, refers to a learning system that uses a linear function to hypothesize space in a high-dimensional feature space. SVM is a supervised learning algorithm in machine learning. In this type of algorithm, a sample set and its corresponding classification identifier are provided for the learning machine. SVM constructs a hyperplane that separates the two classes. During construction, the SVM algorithm tries to maximize the separation between the two classes. The support vector machine is suitable for solving the binary classification problem. The classification principle of SVM is to find a hyperplane that can achieve the optimal classification effect under the premise of the lowest misjudgment rate under the condition of linear separability, which is called the optimal hyperplane. For a set of linearly separable text vectors, the SVM method of linearly separable problems can be used to solve them.

Set the training sample to $(x_1, y_1), \cdots, (x_n, y_n), x_i \in R^n$, $y_i \in \{-1,+1\}, i = 1, \cdots, n$. In the linearly separable case, there

exists a hyperplane $w \cdot x + b = 0$ that can completely separate the two classes of samples, which is ideal. If the hyperplane divides the vector set correctly and the distance sum between these vectors and the hyperplane is the largest, then the hyperplane vector is called the optimal hyperplane, as shown in

$$\begin{cases} (w \cdot x_i) + b \geq 1, & \text{if} y_i = 1, \\ (w \cdot x_i) + b \leq -1, & \text{if} y_i = -1. \end{cases} \tag{1}$$

The rounding formula is shown in

$$y_i[(w \cdot x_i) + b - 1], \quad i = 1, \cdots, m. \tag{2}$$

The optimal plane satisfies the condition as shown in

$$\Phi(w) = \frac{1}{2}\|w\|^2. \tag{3}$$

The training sample point with the smallest hyperplane value is called the support vector, and the classification model of the support vector machine is shown in Figure 1.

In the training sample set, only the k texts that are closest to the new text are calculated, and to determine which category the k texts belong to, the new text will be judged as this category, and the training text vector will be marked according to the feature item set. After the new text to be classified is determined, the new text is segmented according to the feature set, so as to determine the vector representation of the new text. Select the k text vectors that are most similar to the text to be classified in the training text set, and the calculation formula is shown in

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times w_{ik}}{\sqrt{\left(\sum_{k=1}^{M} W_{ik}^2 + W_{ik}^2\right)}}. \tag{4}$$

Among them, there is no general method for how to determine the selection of the k value. Generally, an initial value is preset first, and then, the k value is adjusted according to the result of the experiment. Usually, the initial value is set to a value between three digits and four digits. Calculate the weight of each category in turn, and the calculation formula is shown in

$$p(x, C) = \sum_{d \in KNN} \text{Sim}(x, d_i). \tag{5}$$

The weight of each class is compared, and the text to be classified is divided into which category. Other independent semantic descriptions, in addition to the above independent semantics, are defined as other independent semantics. Under this description, the similarity of meaning items can be defined.

The fundamental reason for modifying the original expression is that the similarity value of the first independent sememe will directly affect the similarity values of other parts. In short, if the similarity of the first independent sememe is not very high, the role played by the proximity of other parts will also be diminished. Based on the idea that the overall similarity can be weighted and calculated by partial similarity, combined with knowledge and experience, effective measures should be taken to ensure that the similarity between the first independent sememe and other sememes does not depend on each other, so that the similarity between the two real words can be calculated. For the selection of weight parameters involved in the calculation process, the biggest advantage of this method is that it can prevent the obtained similarity values from being too similar and can ensure that the obtained calculated values are within the required range.

3.2. Tendency Analysis of Semantic Texts of China-Related Public Opinion. The research on the evolution of China-related public opinion mainly includes two points: one is to analyze the early warning of China-related public opinion crisis, first to determine the important factors affecting the evolution of China-related public opinion and secondly to create a scientific and reasonable index system based on this. Then, the evolution process and trend of China-related public opinion are predicted through this index system; the second is to conduct an in-depth analysis of the evolution model of China-related public opinion, and a detailed analysis of the evolution state of China-related public opinion is carried out according to the interactive activities carried out by individuals. In the final analysis, the analysis of public opinion tendency can be said to be the analysis of text, so the analysis of Chinese public opinion tendency in this paper is based on text. Text orientation analysis [7] refers to the automatic analysis of the positive and negative factors contained in text information through computer technology, occasionally including factors such as positive or negative and like or dislike. It is a subset of text sentiment analysis tasks. The range of emotional factors in text sentiment analysis research is wider, including not only the extreme factors of praise and criticism but also many emotional factors such as happiness or sadness, anger, or fright. The process of text classification is shown in Figure 2.

Since text mining usually deals with large amounts of data, there is a trade-off between the efficiency and accuracy of the model. At present, the mainstream text mining and analysis methods in the industry are mainly based on word frequency, but many cutting-edge analysis methods can already consider the position information of words and the probability of each other. Both the hidden Markov model and the conditional random field can be used for Chinese word segmentation. The hidden Markov model is a widely used method. It has been able to achieve high-precision word segmentation while taking into account the efficiency. The results of Chinese word segmentation can be directly used to build text objects. This paper chooses to build a term-document relationship matrix. This matrix can obtain different results through different text mining algorithms.

Before building a more complex thematic document content model, it is necessary to use the hierarchical clustering method to analyze the similarity between the terms. In order to find out what categories of comments of Weibo users are, we first analyze whether there is a similar

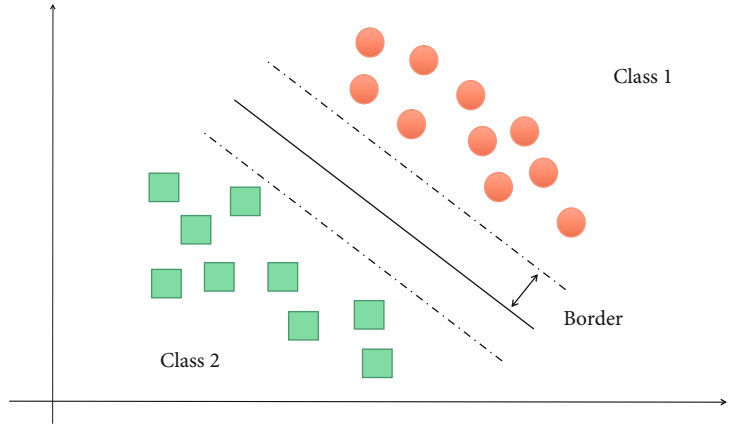

Figure 1: Support vector machine classification model.

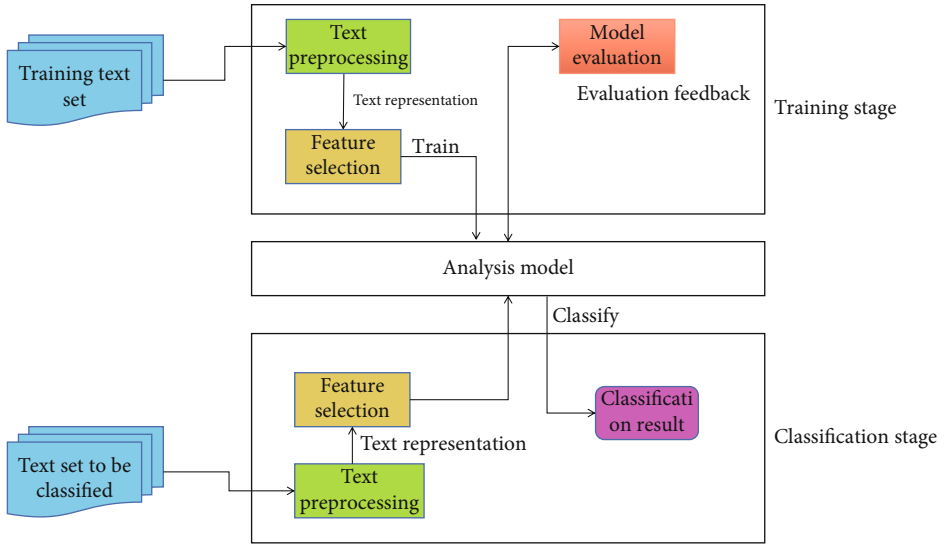

Figure 2: The process of text classification.

relationship between documents and cluster similar texts on the basis of bilateral distance. This method relies on the frequency of terms. The calculation formula is shown in

$$D(G_P, G_q) = \max\left\{d_{ij} \middle| i \in G_p, j \in G_p, p \neq q\right\}. \quad (6)$$

Among them, d_{ij} selects the Euclidean distance as shown in

$$d_{ij}\text{-}\sqrt{\sum(x_{ik} - x_{jk})^2}. \quad (7)$$

The formula for calculating the distance from other classes according to the longest distance method is shown in

$$D(G_r, G_k) = \max\left\{D(G_P, G_k), D(G_q, G_k)\right\}. \quad (8)$$

Different from other text search methods, this method extracts adjacent random samples with some randomness at each step of the search process, continuously searches for the local optimal solution, and takes the best solution as the output result. China-related public opinion information caused by emergencies is mainly text in terms of content and form. Therefore, the classification of public opinion information is essentially text classification, and the existing research results related to text classification can be applied to the classification of public opinion information. Supervised and semisupervised learning on imperfect data machines has been studied. Traditional data mining and machine learning algorithms make predictions on unknown data using statistical models trained on previously collected labeled and unlabeled training data. Semi-supervised classifiers solve the problem of too little labeled data to build a good classifier by using a large amount of unlabeled data and a small amount of labeled data.

4. Experimental Results and Performance Analysis

4.1. Data Search and Evaluation Indicators. Text mining can be subdivided into various types according to different standards and requirements. For example, it can be subdivided

into two types according to the classification of document objects: one is text mining based on a single document. This method is only applicable to a single document. The in-depth study of a single document obtains tacit knowledge, and its mining technologies mainly include information extraction technology, document summarization technology, and so on. The second is the mining method based on the document set. This method can extract patterns from a large number of document data. The mining technology mainly includes document filtering and text clustering.

In order to verify the automatic extraction performance of the method in this section, this paper selects 100 forum sites as the data source, including some representative Chinese forums and some English forums. Some pages are selected as experimental data for the determination of parameters in the method, and the remaining pages are used as test sets to verify the performance of the method. The data composition is shown in Table 1.

The main parameters used in the method are the layout similarity threshold, the number of layers to be compared when calculating the layout similarity, and the contribution coefficient of each layer to the overall layout. Among them, the selection of the number of comparison layers N is the most important, because it affects the selection of other parameters, and the following factors must be considered: if the number of layers is too small, that is, if the comparison is too rough, many similar nodes may be obtained, which affects the accuracy of the final extraction. If too many layers exist, that is, if the comparison is too detailed, you may not get the correct result and consume more time.

In the experiment, different N is selected to calculate the average time of processing each web page and the recall rate and accuracy rate of extraction. The value of N ranges from 1 to 8. The experimental results are shown in Figure 3.

Figure 4 shows that the processing time increases with the increase in N and eventually tends to be flat. This is because most of the DOM tree layers of web pages are within a certain range, so the actual number of layers processed will not vary with N. Comprehensive consideration makes the extraction accuracy and recall rate relatively high, and the operation processing speed is fast. The selection of the contribution coefficient should be based on the following principles: the contribution coefficient of the outer layer is greater than that of the inner layer, and the experimental analysis of the accuracy and recall rate of the processing is carried out, and the experimental results are shown in Figure 4.

It can be clearly seen from Figure 5 that when N is too small or too large, neither the precision nor the recall can achieve satisfactory results, which is consistent with our analysis. Considering it comprehensively, N generally takes 2 or 3; at this time, a relatively balanced state can be achieved in terms of processing speed and extraction accuracy.

In the experiment, this paper chooses the value of N to range from 1 to 1 to calculate the average time for processing each web page and the recall rate and accuracy rate of extraction, which can be adjusted according to the actual situation. With the group reference value, the effect is better when the layout similarity is 0.9. The specific experimental results are shown in Figures 3 and 4. Figure 3 shows that

Table 1: Experimental data and test data.

	Number of test pages	Number of experimental web pages
Main page	7382	218
Content page	15622	294
Total	23004	512

the processing time increases with the increase in N and eventually tends to be flat. This is because most of the DOM tree layers of web pages are within a certain range, so the actual number of layers processed will not change with the increase in N. It can be clearly seen from Figure 4 that when N is too small or too large, neither the precision nor the recall can achieve satisfactory results, which is consistent with our analysis. Considering it comprehensively, N generally takes 2 or 3; at this time, a relatively balanced state can be achieved in terms of processing speed and extraction accuracy.

4.2. Analysis of Chinese-Related Public Opinion Tendency Based on Corpus. The discourse analysis of China-related public opinion tends to focus on qualitative research. This paper argues that critical discourse analysis often selects specific texts for research and lacks objectivity. Moreover, it chooses to analyze those features of the discourse that support its point of view, so its interpretation is biased. In response to this question, some scholars have introduced corpus research methods into critical discourse analysis. Quantitative analysis provides a solid data foundation for corpus research, which helps to overcome the subjectivity and one-sidedness of researchers; the richness of corpus also reduces the randomness of researchers in choosing analysis objects and enhances the persuasiveness of explanations. The method in this paper is used to retrieve the 5 collocations before and after the word corruption, and the words with the most practical meanings are selected for research, as shown in Table 2.

The high-frequency words and relevance information provided by the vocabulary can reflect the overall characteristics of the text, but mining the implicit attitude and ideology of the text requires further observation of the context in which the keywords are located. These terms reflect the Western media's recognition of China's anticorruption efforts. Its basic idea is that when only partial knowledge about the unknown distribution is obtained, the probability distribution model that conforms to this knowledge and has the largest entropy should be selected. Because there may be more than one probability distribution that conforms to known knowledge and entropy is the amount of information that reflects the uncertainty of random variables, finding the probability distribution that maximizes entropy is the most uncertain inference for conforming to known knowledge. Use a maximum entropy classifier and separate subjective and objective texts.

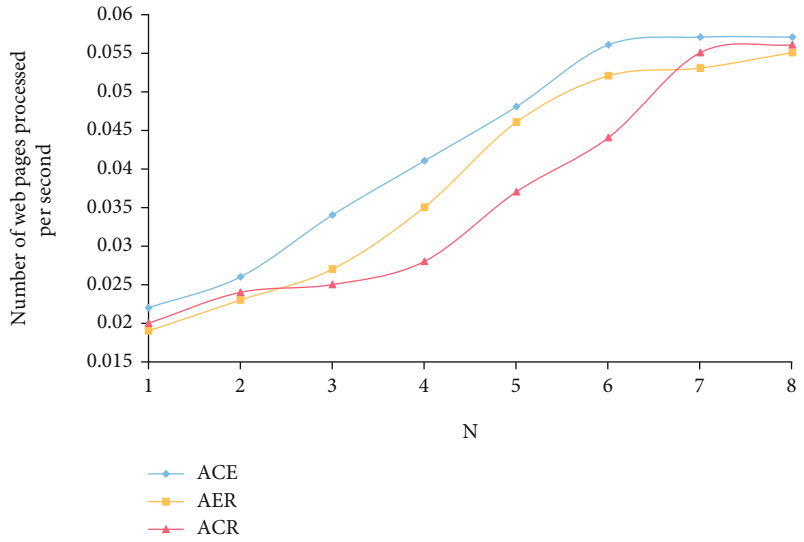

Figure 3: Average processing time for different values of N.

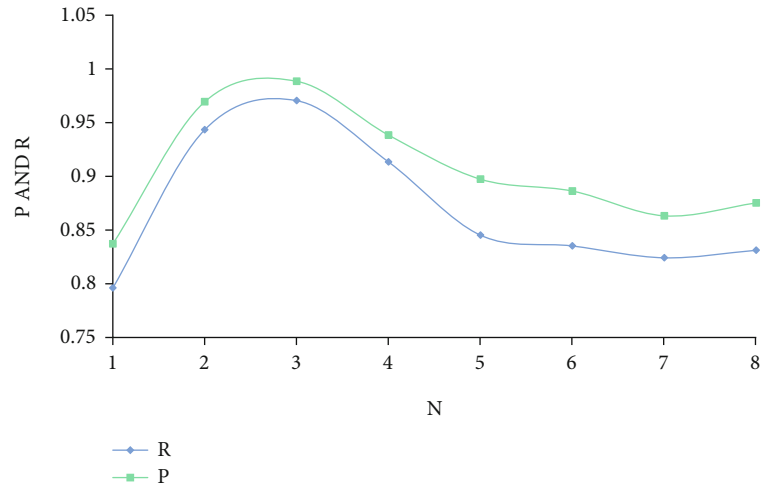

Figure 4: Recall and precision when N takes different values.

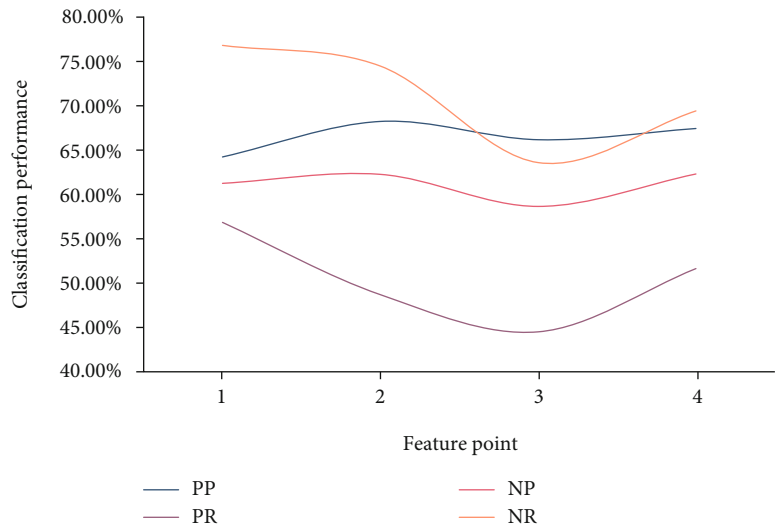

Figure 5: Comparison of classification performance of N-grams.

Table 2: High-frequency collocation words with corruption.

Serial number	Collocation words	Word frequency	Word frequency of collocations on the left	Word frequency of collocations on the right
1	Anti	94	93	2
2	Campaign	41	6	36
3	Against	25	22	4
4	Drive	23	0	23
5	Official	23	21	3
6	President	23	18	6
7	Party	22	11	12
8	Government	21	10	12
9	Chinese	20	8	12
10	Crackdown	15	10	6

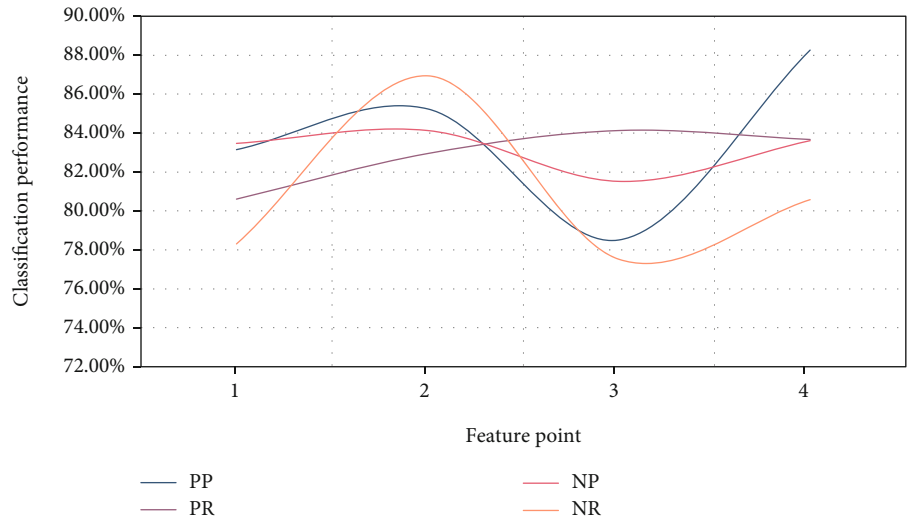

Figure 6: Comparison of classification performance of PMML+N-gram using the NB classifier.

4.3. PMML Experiment Based on the Method of Matching and Marrying Machine Learning. In the process of selecting classifiers and features, this paper first adopts the *N*-gram feature selection method, selects the SVM classifier for classification experiments, and then uses the combination of pattern matching and statistical learning proposed in this paper. Classifiers, Bayesian methods, and SVM methods were compared with different combinations of *N*-grams to conduct experiments. Before this kind of technology extracts information, users need to manually design information extraction rules for each website, and then, the information extraction system extracts information according to these rules. Although the current automatic information extraction technology no longer requires manual participation, it can be applied to the acquisition of massive information. However, due to the lack of granularity and semantic classification of the extracted information, this type of extraction technology is not suitable for the acquisition of data sources for China-related public opinion analysis. However, in the actual situation of Chinese-related public opinion research, there are a lot of ambiguous phenomena in the definition of set attribution, and the natural language used to describe the concept has no clear extension, and it is a fuzzy concept. In order to make effective use of the fuzzy expression of the evaluation object, it can be fuzzed and expressed as a triangular fuzzy value, and the output result contains more information through the fuzzy weighted average, which provides rule basis and data cleaning for China-related public opinion news.

In this experiment, three types of feature representation methods are used, namely, unigrams, bigrams, and trigrams. 70% of the review data is used as the training set, and the remaining 30% of the review data is used as the test set. All features are selected, and the SVM classification method is used. The SVM classifier adopts MATLAB7.0 and SVM Toolbox. The experiments in this paper are all carried out after its modification. The kernel function selects the sigmoid function, where $a = 1$ and $b = 1$. The performance evaluation indicators use PP, PR, NP, and NR to represent positive forward precision, forward recall, reverse precision, and reverse recall. The experimental results are shown in Figure 5.

As shown in the classification results in Figure 5, the experimental results are unsatisfactory in terms of accuracy

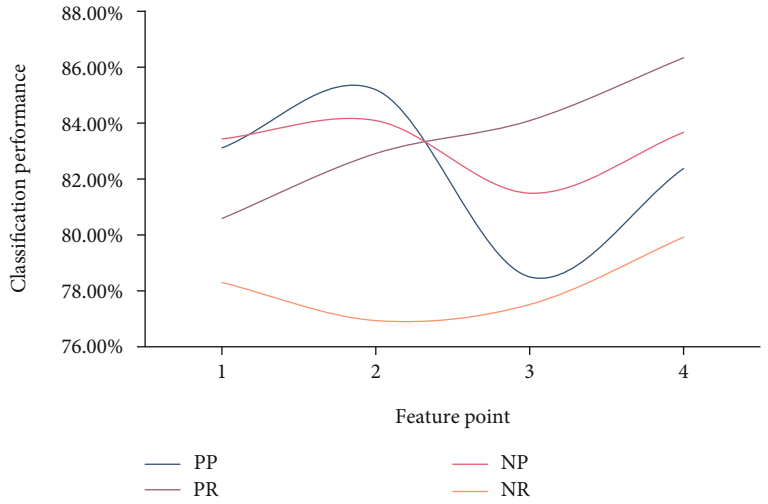

Figure 7: Comparison of classification performance of PMML+N-gram using the SVM classifier.

and recall and are quite different from the performance results of text classification. This has a lot to do with the fact that the corpus text is directly selected from the Internet, the language is not standardized, the Internet vocabulary exists, and the styles are different. Overall, the experimental results produced by each feature representation method in N-gram bigrams are slightly better than the other two.

Figures 6 and 7 show the results of NB and SVM classification, respectively, after pattern matching the text.

The experimental results show that the classification using the bigram method as a feature item is still slightly better, and the best accuracy rate in the SVM experiment is 86.42%. By analyzing the reasons, this is mainly because words with positive or derogatory tendencies appear in a sentence or article, which basically determines the semantic tendency of the sentence or article. In the pattern matching, the window length of the emotional pattern feature is selected as 3, but in many cases, 2 words already constitute the emotional feature.

The experimental results show that the effect of the method proposed in this paper is improved by about 10% compared with the direct use of statistical learning methods for emotional orientation analysis. The SVM method also has obvious advantages over the Bayesian method, especially the combination of the SVM method and the bigram method which gives the best results. As analyzed in this paper, if all words are regarded as independent individuals, the dependencies between the feature items in the text will be lost, but the method in this paper removes some words that are not related to the tendency after selecting the pattern and calculated the inclination of keywords and patterns, taking into account the relationship between the features in the text and the inclination of words and then using machine learning methods to naturally improve the performance of classification.

5. Conclusions

This chapter introduces and analyzes sentiment orientation analysis and introduces different granularity sentiment orientation analysis methods, including word-level, sentence-level, and text-level orientation classification methods. By introducing the inadequacies of dictionary-based and statistic-based methods, this chapter proposes a method based on pattern matching and statistical learning to analyze the sentiment tendency of web comments; first segment the text and then extract according to keyword categories and then according to Chinese sentiment the contextual expression habit, and the methods provided in other literatures give ten types of patterns, and the extracted keywords are matched according to these ten types of patterns. Statistical learning methods, such as Bayesian method and SVM method, are used to obtain the final propensity on the sentiment pattern vector. The experimental results show that the effect of the method proposed in this paper is improved by about 10% compared with the direct use of statistical learning methods for emotional orientation analysis. The SVM method also has obvious advantages over the Bayesian method, especially the combination of the SVM method and the bigram method which gives the best results. However, we should also see that the corpus used in this experiment is selectively obtained from the Internet, and its attitude is relatively clear and the words are relatively standardized. In order to be able to better classify the tendency of web comments, it is necessary to analyze the expression of Chinese sentiment in more detail and effectively expand the pattern library.

The mining and analysis of China-related public opinion information is a research field full of opportunities and challenges. It also involves the common development of multiple disciplines. There are still many issues to be further explored and studied. Text is the main carrier of Chinese-related public opinion information. How to effectively discover the semantic information contained in the text is one of the main factors restricting public opinion mining. The method of local latent semantic analysis in this paper mines the semantic relationship from the internal relationship between the uses of feature words, which is also a relatively shallow semantic mining. The future work will be to study the transformation of features into semantic concepts through

semantic understanding and calculation, so as to retrieve information related to this concept and overcome the limitations of traditional information retrieval technology.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The author declares that he has no conflicts of interest.

Acknowledgments

This work was supported by the Project of Humanities and Social Sciences from the Ministry of Education of the People's Republic of China (21YJAZH125): "Research on the corpus-based Critical Discourse Study of China-related Public Opinion from the Five Eyes Alliance".

References

[1] J. Du and J. Cai, "Analysis of data acquisition and text classification technology for network public opinion monitoring," *Wireless Internet Technology*, vol. 4, no. 15, pp. 232–241, 2019.

[2] W. Zhang, X. Li, H. He, and X. Wang, "Identifying network public opinion leaders based on Markov logic networks," *The Scientific World Journal*, vol. 2014, Article ID 268592, 8 pages, 2014.

[3] L. Wright, A. Burton, A. Mckinlay, A. Steptoe, and D. Fancourt, "Public opinion about the UK government during COVID-19 and implications for public health: a topic modeling analysis of open-ended survey response data," *PLoS One*, vol. 17, no. 4, p. e0264134, 2022.

[4] D. Ruiter, L. Reiners, A. G. D'Sa et al., "Placing M-Phasis on the plurality of hate: a feature-based corpus of hate online," 2022, https://arxiv.org/abs/2204.13400.

[5] Y. Yisimayili, T. Yibulayin, and K. Abiderexiti, "Research of user-relationship based data acquisition method on Uyghur microblog," *Journal of Xinjiang University (Natural Science Edition)*, vol. 5, no. 1, pp. 46–55, 2015.

[6] M. J. Zhang, "Design and implementation of network public opinion data acquisition system based on web spider," *Modern Computer*, vol. 3, no. 1, pp. 70–80, 2015.

[7] S. U. Peng and W. S. Yang, "Monitoring the Internet public opinion and promoting intellectualization of CPC construction in ethnic minority areas of China," *Journal of Yunnan Minzu University (Social Sciences)*, vol. 6, no. 1, pp. 32–45, 2019.

[8] W. D. Yin, X. H. Zhu, and J. K. Zhao, "Technical analysis of Internet public opinion," *Netinfo Security*, vol. 3, no. 1, pp. 62–77, 2012.

[9] Y. Li, Y. He, M. Cai, Y. Zheng, and X. Tan, "Comparative analysis and effective strategies of government response to network public opinion," *Journal of Intelligence*, vol. 8, no. 1, pp. 91–101, 2018.

[10] F. Tian and J. Shen, "Microblog data extraction algorithm based on user influence," *Computer Applications and Software*, vol. 8, no. 2, pp. 26–38, 2017.

[11] L. Xue-Feng and S. Y. Chen, "Review of natural disaster network public opinion information analysis and management," *Geography and Geo-Information Science*, vol. 3, no. 3, pp. 43–52, 2016.

[12] H. Wang and G. Wang, "Government intervention mechanism of social public opinion crisis," *Bulletin of the Chinese Academy of Sciences*, vol. 2, no. 1, pp. 73–78, 2015.

[13] L. Cui, "The overall design of the intelligent analysis system of network public opinion," *Journal of Henan Institute of Science and Technology (Natural Science Edition)*, vol. 2, no. 1, pp. 6–14, 2015.

[14] Q. Y. Hu and M. Hong, "Acquisition and technology of Internet public opinion analysis," *Journal of Gansu Police Vocational College*, vol. 7, no. 1, pp. 33–40, 2014.

[15] M. Wang and J. Sun, "Generation mechanism of corporate online public opinion hotness based on multicase qualitative comparative analysis," *Discrete Dynamics in Nature and Society*, vol. 2021, 11 pages, 2021.

[16] M. J. Ha, "Corpus-based literary analysis," *Journal of the Korea Safety Management & Science*, vol. 13, no. 9, pp. 440–447, 2013.

[17] H. Jeong, S. Shiramatsu, K. Kobayashi, and T. Hatori, "Discourse analysis of public debates using corpus linguistic methodologies," *Journal of Computers*, vol. 3, no. 8, pp. 58–68, 2008.

[18] S. Fan, "The analysis of synonyms Based on COCA corpus–taking "speak" and "tell" as an example," *Campus English*, vol. 5, no. 1, pp. 46–52, 2017.

[19] C. Q. Liu, Y. Xu, and J. F. Xu, "The content acquisition method of Libnids-based university network public opinion," *Applied Mechanics & Materials*, vol. 63-64, pp. 775–779, 2011.

[20] Y. Chen and Y. Huang, "Dynamic analysis of online public opinion," *Information and Documentation Services*, vol. 4, no. 4, pp. 76–79, 2016.

[21] L. U. Jia, "Data analysis and expectation of corpus-based translation studies in China (2008-2018)," *Journal of University of South China (Social Science Edition)*, vol. 7, no. 13, pp. 62–68, 2018.

[22] L. N. Shen, "Corpus-based acquisition features and error analysis of Chinese company adverb," *Journal of Tangshan Normal University*, vol. 5, no. 18, pp. 84–89, 2018.

[23] P. Wang, H. Xue, and F. Zhang, "Feature representation and organization method for public opinion big data based on association analysis," *Journal of Physics: Conference Series*, vol. 1881, no. 3, article 032075, 2021.