Hindawi
Scientific Programming
Volume 2021, Article ID 9347432, 11 pages
https://doi.org/10.1155/2021/9347432

Research Article

A Study on the Relationship between Public Derivative Big Data and Industrial Policymaking: Taking Bike Sharing as an Example

Huilin Song

School of International Economics and Trade, Jiangxi University of Finance and Economics, Nanchang 330013, China

Correspondence should be addressed to Huilin Song; songhuilin@jxufe.edu.cn

Received 6 May 2021; Accepted 3 July 2021; Published 16 July 2021

Academic Editor: Yi-Zhang Jiang

Smart government is an important part of the smart world. The use of big data analysis technology can effectively improve the government's ability of fine management. Taking China's bike-sharing industry as the research object, we study the relationship between public-derived big data and industrial policy. First, a feature-enhanced short text clustering method is proposed to perform topic clustering on publicly derived big data. Second, keyword extraction based on word frequency is used to quantify the text of industrial policy. Finally, time is taken as the main line to analyze the co-occurrence of clustering topics and keywords. The results show that (1) the feature enhancement method we proposed can effectively improve the clustering effect. (2) There is a great correlation between the industrial policy and the information mined by Weibo, but there is an obvious lag. Rational use of public-derived big data will effectively help the industrial policy to be released in a better and faster way.

1. Introduction

With the development of mobile communication technology, the mobile Internet, and the popularization of mobile smart devices, the characteristics of the widespread of mobile Internet information and the large coverage have made people enter an era of information explosion. Since the era of "We-Media," the impact of public-generated big data [1] on public policy changes has become more pronounced [2]. Among the many ways of publishing information on the Internet, Weibo (Chinese microblog) has the characteristics of limiting the total number of input words (no more than 140 words), as well as strong interactivity and time-sensitive, makes the focus information in the text easier to be highlighted. It has gradually become an important information channel for the general public to discover news, release information, and discuss. In China, almost all mainstream official media and government propaganda departments at all levels have registered Weibo accounts. These official Weibo accounts of the government or media have high updating frequency and strong interaction, which have become a new platform for communication with the public and a channel for direct dialogue between the public and policymakers [3].

The sharing economy is based on the Internet as a platform to provide individuals or organizations with idle resources to temporarily transfer the right to use the service. Taking bike sharing as an example, it effectively improves the utilization efficiency of social resources, optimizes the allocation of resources, solves the "last kilometer" traffic pain point of residents, and meets the needs of different groups. However, driven by the brutal and disorderly capital, bicycle-sharing companies have the uncontrolled release of shared bicycles in cities to quickly seize the market, resulting in oversupply. This behavior not only violates the original intention of the "sharing economy" to optimize the allocation of resources but also brings new social problems. The disorderly parking of bicycles, the disposal of old bicycles, and other problems not only bring trouble to the city management but also affect the normal traffic order. Excessive competition leads to poor operation of enterprises, and some enterprises even delay refund or withhold user deposits. This series of problems exposed the lag in the government's policy formulation for the development of the new industrial norms of the bike-sharing industry. Many users of the above-mentioned problems have posted on Weibo for the first time, but due to the fragmented nature of Weibo information, it is not easy to collect, organize, and mine this

information. Therefore, how to mine the effective information of shared bicycles in Weibo, analyze its relationship with industrial policies, and finally provide decision support for the government to formulate corresponding industrial policies is the focus of this paper.

The overall research framework of this paper is shown in Figure 1. Firstly, relevant data are collected through web spider, then topic extraction and classification of Weibo data is carried out through the BTM topic model, keyword extraction based on word frequency is carried out on policy text, and finally, co-occurrence of problems and policies is analyzed based on time series, to provide strong decision support for the formulation of industrial policies.

2. Related Work

2.1. The Public Derivative Big Data and Public Policy. All social activities will generate data, and various technologies for analyzing massive data have emerged. How to use data and data analysis technology to optimize public policy issues and improve the quality of decision making has become a research hotspot. In the field of management, big data analysis technology is the first to be applied to the business field by large multinational companies [4]; through the mining of multidimensional data, it provides effective support for the final decision. With the application of this technology in public government affairs, research on the relationship between public management, public policy and public-derived big data has gradually emerged [6–9]. Public-derived big data refer to the data released by non-professionals around public topics and public affairs on the Internet platform. These data are characterized by complexity, diversity, and low-value density.

Cai and Yang [10] put forward an application framework in a big data environment that includes four parts: the construction of past case base, the detection and analysis of current social public opinion, the early warning of future public opinion, and the support of public policy decision making. Different computational models have been applied to the mining of commonly derived big data, including text mining, semantic understanding, sentiment analysis, and hot spot discovery [11]. Ma et al. [12] introduced the topic discovery model LDA in the analysis of public messages on the interactive platform of Chaoyang District, Beijing. It first measures the cost elements, then quantifies the service efficiency, and discusses the relationship between the public service efficiency and the cost. Chun et al. [13] used the Bayesian classification model to analyze the public environmental policies of different races under the condition that income and political behavior and other factors remained unchanged. Li et al. [14] improved the traditional MapReduce method and mined the relationship between text feature vectors and the development of public opinion.

2.2. The Topic Discovery Model for Short Texts. Text topic discovery has always been a focus in the field of natural language processing. The bag of words (BOW) model was first used in the discovery of text topics. The model regards each document as a combination of multiple words and assumes that the relationship between words is independent [15]. Meanwhile, the default words in the document are out of order, ignoring the grammar and word order of the text of the document and simply using words to represent the topic of the document. This unordered method was gradually replaced by the LDA (latent Dirichlet allocation) [16] topic model due to poor performance. Based on Twitter, Weng et al. [17] considered combining a user's tweets into a single document and then used LDA for training. Similarly, Honey and Herring [18] took into account that the microblogs published by different people may involve many different topics, so they used Twitter messages containing the same words for aggregation. However, such a method greatly depends on the validity of the dataset, and the effect after aggregation cannot be guaranteed. After the BTM [19] model was proposed, it showed good performance not only in long text processing but also in short text processing and was considered to be a good substitute for the traditional topic model of LDA. Tang [20] used BTM model to represent microblog feature vectors and combined the traditional vector space model (VSM) with BTM according to certain weights to make up for the deficiency of VSM. Zhang et al. [21] studied the use of BTM's topic-vocabulary matrix as the external knowledge to expand the VSM vector, effectively solving the problem of microblog data sparseness.

3. Feature Reinforcement Topic Model for Short Text

3.1. Feature Vector Enhancement Method. The multimeaning or synonym of a word is the key to semantic understanding, especially in the task of text topic discovery in a specific domain. Therefore, in this task, the model performance is often poor because of the special definition of words in a certain domain. For example, the word "WeChat" (a mobile instant messaging app that can be used for payment) may appear for two different topics of social interaction and payment, so social interaction and payment are the potential topics of "WeChat," and we need to strengthen the features, respectively. The enhancement method adopted in this paper is VSM. Figure 2 shows the feature enhancement method.

Formally, for a document D that needs to discover a topic, it may contain feature f, which contains the following three cases:

(1) When feature f only belongs to the topic T_i, T_i is used to identify feature f during reinforcement, and the weight of the topic T_i is denoted as the weight of feature f.

(2) When the potential topic of feature f includes T_{i_1}, T_{i_2},. . .. . ., T_{i_m}, m topics need to identify the characteristics of the current document at the same time.

(3) When feature f does not belong to any topic T, this feature is not reinforced, and f itself is used to identify the document feature.

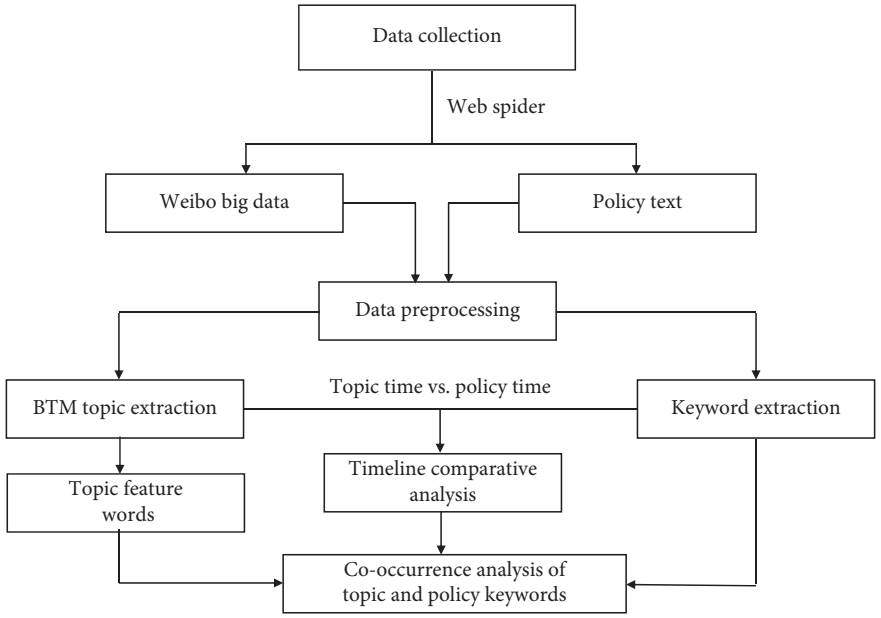

Figure 1: A framework of research on the relationship between public-derived big data and policymaking.

Word	WeChat	Photo	Envelope	Voice	Moments	5 yuan	Can	……
Weights	0.20	0.13	0.22	0.12	0.11	0.19	0.04	……

Vanilla VSM

Weibo feature vector representation:

Weibo A = (WeChat w_1, Photo w_2, Envelope w_3, Voice w_4, Moments w_5, 5 yuan w_6, Can w_7)

Weibo B = ……

Weibo C = ……

Weibo D = ……

Cluster 1 Cluster 2 Others

Feature enhancement

Clustering results:

Cluster 1 (social) = {Weibo A, Weibo D, ……} description : WeChat, photos, envelopes, voice, moment……

Cluster 2 (pay) = {Weibo B, Weibo C, ……} description : WeChat, envelopes, 5 yuan……

Enhanced Weibo A = (feature1 : f_1, feature2 : f_2, Can : w_8)

Where $f_1 = a_1w_1 + w_2 + b_1w_3 + b_1w_4 + w_5, f_2 = a_2w_1 + b_2w_3 + c_2w_6$

- The coefficients a, b, and c are determined by the distribution probability of feature words under different topics

After feature enhancement

Figure 2: An example of feature enhancement.

For vector V, it is assumed that f_1, f_2, f_3, and f_4 belong to topic T_1, f_4 and f_5 belong to topic T_2, and f_6 and f_7 do not belong to any topic. V' is the Weibo vector after feature enhancement. Figure 3 shows the process of feature enhancement.

There are three advantages after feature combination: (1) T_1, the main potential topic of the text, is highlighted; (2) selective reinforcement also takes into account the possibility of f_4 becoming a secondary potential topic T_2; and (3) for general terms with no topic such as f_6 and f_7, their sparsity can be kept as suppression. The enhanced VSM vector, to some extent, reduces the dimension and highlights the main potential topic of the document.

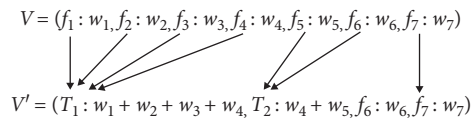

Figure 3: Formal process of feature enhancement of VSM vector.

When the fixed vector dimension M is maintained, the enhanced vector can retain more original information of the document, highlight the main features, and better describe the document.

3.2. BTM Model. BTM is a generation model, and its generation process is shown in Figure 4. The BTM model consists of three layers, namely, word pair, topic, and word, in which word pair to topic obeys Dirichlet distribution and topic to word obeys polynomial distribution. Considering that the entire document is a mixture of multiple topics, the BTM model algorithm can further alleviate the problem of data sparsity and facilitate the topic discovery of short texts by learning from the global topic library.

Specifically, the algorithm process of the BTM model can be described as follows.

Firstly, the word distribution under a single topic z is plotted, i.e., $\varphi_z \sim \mathrm{Dir}(\beta)$, where $\mathrm{Dir}(\beta)$ represents the Dirichlet distribution of words.

Secondly, the global topic distribution of the whole document is plotted, i.e., $\theta \sim \mathrm{Dir}(\alpha)$, where $\mathrm{Dir}(\alpha)$ represents the polynomial distribution of the topic.

Next, operate on b for each word, assuming $b = (w_i, w_j)$.

(a) Select a topic z from the global topic distribution θ, i.e., $z \sim \mathrm{Multi}(\theta)$.

(b) Choose two words w_i and w_j from topic z, i.e., w_i, $w_j \sim \mathrm{Multi}(\varphi_z)$.

According to the above process, the joint probability $P(b)$ of the word against b can be calculated by the following formula:

$$P(b) = \sum_z P(z)P(w_i|z)P\left(w_j|z\right) = \sum_z \theta_z \varphi_{i|z} \varphi_{j|z}. \quad (1)$$

Among them, $P(w_i|z)$, $P(w_j|z)$, respectively, represent the probability that words w_i and w_j belong to topic z. Further, the probability of the entire document topic is calculated by the following formula:

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \varphi_{i|z} \varphi_{j|z}. \quad (2)$$

3.3. BTM Model Based on Feature Enhancement. The generation of VSM vectors by Weibo text requires the selection of appropriate terms as vector dimensions. The traditional document frequency method based on statistics can only remove noise through screening and cannot solve the problem of feature sparsity caused by the small number of words in Weibo. For this reason, we propose a feature enhancement method based on BTM, the process of which is shown in Figure 5.

In the process of BTM modeling, we construct the feature distribution matrix for the topic and then use the topic-lexical matrix to consolidate the vector features.

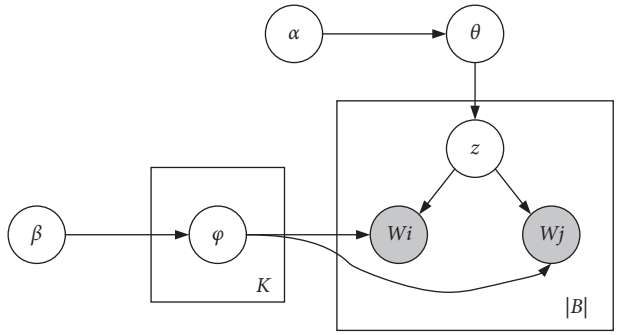

Figure 4: BTM model generation process diagram.

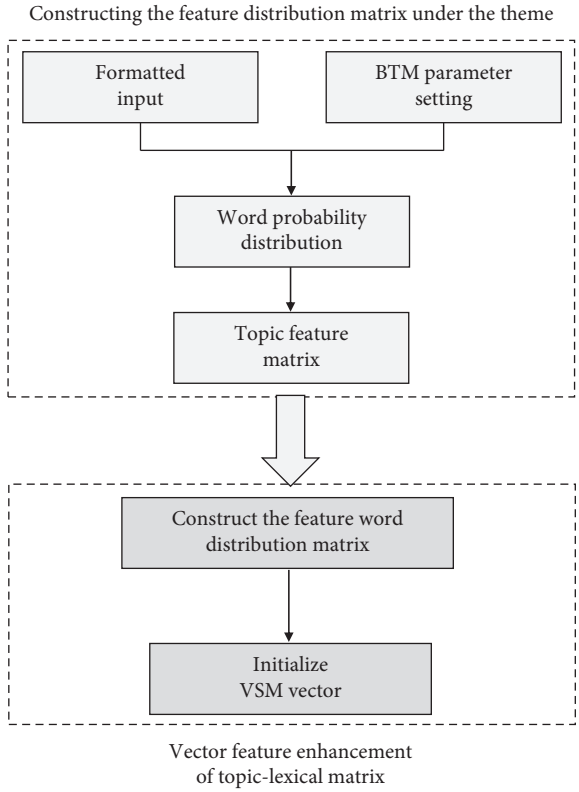

Figure 5: Flowchart of feature enhancement based on BTM.

3.3.1. Constructing the Feature Distribution Matrix under the Topic

Step 1: format the input. Input the VSM vector of each behavior in the aggregated document, and the VSM vector is generated after the selection of Weibo short text participles, stoppages, and features. First, we generate the dictionary text from all the words after segmentation, then map the VSM vector with the dictionary text, and finally get the formatted input text.

Step 2: parameter setting. In the BTM modeling in this paper, the parameter α is $50/K$, where K is the number of topics, and different values are set for different tasks. The

parameter β was 0.01; the number of iterations is set to 1000.

Step 3: generation of the word probability distribution. The words under each topic are arranged in descending order according to their distribution probability and spliced into the format of "words: weight," which is stored in the probability text file. One line represents the distribution probability of all words under a topic.

Step 4: generate the topic-feature matrix. The words with a probability higher than 0.001 in the probability text were selected as candidate words and used as the feature words of the topic, and ID mapping was carried out to generate the topic-feature matrix.

3.3.2. Vector Feature Enhancement of Topic-Lexical Matrix. After the topic-feature database is obtained, to facilitate the calculation of the algorithm, we need to reverse calculate the probability distribution of each candidate feature under different topics and carry out normalization processing, to allocate the weight during selective reinforcement. The matrix structure is as follows:

$$\begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0k} \\ p_{10} & p_{10} & \cdots & p_{1k} \\ \cdots & \cdots & & \cdots \\ p_{n0} & p_{n0} & \cdots & p_{nk} \end{bmatrix}, \quad (3)$$

where p_{ij} represents the probability of the i-th word under topic j, i is 0 to n, j is 0 to k, n is the number of candidate feature words, and k is the number of candidate topics. Then, the VSM vector is initialized. In this paper, each weight vector is added k dimensions and initialized to 0, that is, $t_i = 0$. According to the feature vector enhancement method introduced in Section 3.1, for the vector V_{weight}, search in the feature word distribution matrix according to each word in the VSM vector, complete the feature enhancement process of all topics, and update the weight value in V_{weight}'.

4. Empirical Data

4.1. Data Collection. The data time interval selected in this paper is from July 2016 to October 2017, which covers the three market development stages experienced since the birth of shared bikes, which can depict the development trajectory of shared bikes. Figure 6 shows the bike-sharing market and user timeline including three stages of development. The period of the first stage is before October 2016. At this stage, the market is mainly concentrated in developed cities such as Beijing, Shanghai, Guangzhou, and Shenzhen. There are no more than five service providers and the user scale is just over one million. The period of the second stage is from October 2016 to June 2017. At this stage, the market has covered the mainstream first-tier cities in China. The number of service providers has soared to dozens, and the user scale has reached about 50 million after experiencing explosive growth. The period of the third stage is after June 2017. At this stage, the market has begun to sink to third-tier cities. Many service providers have closed down or are forced to merge, and the user scale is stable at about 70 million. To accurately reflect the characteristics of each stage, we named the first stage as the user training stage, the second stage as the demand outbreak stage, and the third stage as the elimination stage.

The data are divided into two parts: Weibo data and policy documents. Since the amount of Weibo data is huge, we use the multithreaded web spiders [22] based on the Scrapy framework for automatic collection. Table 1 shows the details of Weibo data acquisition, including data sources, time range, acquisition tools, keywords, and screening rules.

Due to the limited number of policy documents, we manually downloaded and collected them from the government website.

4.2. Data Preprocessing. Both Weibo and policy documents are in text format, which cannot be directly modeled and analyzed. Therefore, we preprocessed the text data first. Weibo data preprocessing is divided into three steps.

The first step is word segmentation. The word segmentation tool selected in this paper is Jieba word segmentation [23], which provides a custom user dictionary function. To achieve a better word segmentation effect in specific fields, the field words in the Sogou input method lexicon are selected in this paper to improve the effect of high score words.

The second step is to remove emoticons and stop words. First, emoticons are removed from all texts. Then, more than 3,000 discontinued words, including punctuation marks, meaningless numbers, mood particles, appellation words, etc., are selected from the online collection.

The third step is to calculate the weight and calculate the TF-IDF value according to the above-obtained text to form the original VSM vector after preliminary processing.

The policy document data preprocessing is relatively simple, only requiring word segmentation. The word segmentation method is similar to the first step of Weibo data preprocessing, and the method will not be described in detail.

5. Result and Analysis

5.1. Feature Enhancement Results. The initial VSM vector is feature-enhanced using the matrix of potential topic high-frequency words, that is, FE-VSM (feature-enhanced vector space model). In order to verify the effect, the following three kinds of vectors were used as references to conduct comparative experiments: (1) VSM vector without feature enhancement; (2) LDA topic vector; (3) feature vector combined with BTM and VSM. We compare the quality of clustering when the vector dimensions are 100, 300, 500, 700, 900, 1100, 1300, and 1500. The text similarity uses cosine similarity to perform standard K-means clustering. The k value is 12 (BTM optimal number of topics), and the initial center is randomly selected. To eliminate the chance of random selection, the results are averaged 10 times. The experimental results are shown in Table 2.

The results show that the effect of traditional VSM is the worst, and the effect of LDA is not good, and it is also very unstable, which is caused by the inadequacy of LDA in a short text. Combining the vector represented by the BTM

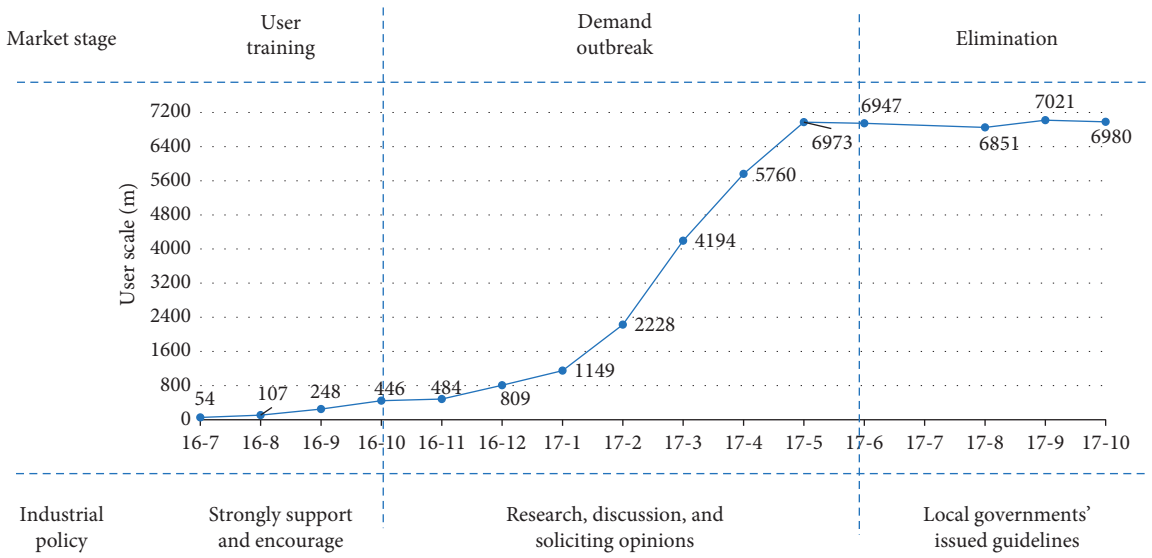

Figure 6: Timeline of the bike-sharing market and user size.

Table 1: Weibo posts' data collection details.

Source	http://www.weibo.com
Time range	From July 2016 to October 2017
Method	Multithreaded web spiders based on Scrapy framework
Keyword	Direct keywords: "shared bike," "Internet bike," etc. Bike brands: Mobike, Ofo, etc. Brand name: "little yellow car," "little blue car," and so on
Filtering rules	Rule 1: keep only original Weibo and exclude reposts Rule 2: keep only one Weibo for repeated Weibo

Table 2: Comparison of results of different topic clustering methods.

Vector dimension	VSM	LDA	BTM + VSM	FE-VSM
100	0.150612	0.18651	**0.304817**	0.244714
300	0.253034	0.185977	0.321333	**0.365615**
500	0.253241	0.211059	0.385545	**0.435322**
700	0.217615	0.266491	**0.386406**	0.355565
900	0.224406	0.150216	0.38029	**0.416342**
1100	0.227279	0.294109	0.373178	**0.383586**
1300	0.240335	0.208641	0.364237	**0.395509**
1500	0.224746	0.261475	**0.369184**	0.366157

The bold values indicate best results.

and the VSM vector according to a certain weight, the effect has been significantly improved. The clustering quality of BTM-enhanced vectors is also significantly better than that of unenhanced vectors, indicating that the method in this article solves the problem of data sparsity and expression diversity to a certain extent, and the enhanced vector can better describe Weibo text information.

5.2. User Training Stage Analysis

5.2.1. Weibo Topic Mining and Analysis. At this stage, the number of Weibo posts was relatively small, including 2,253 Weibo. The topic-feature word distribution results of topic clustering are shown in Table 3. All Weibo are clustered into 4 different topics. Through the analysis of the feature words of topics 1, 2, 3, and 4, it can be found that most of the Weibo clustered by topic 1 is praise content released by users after using shared bikes. We call such topics advantage feedback. Topic-clustering Weibo expresses users' expectations and freshness, including the content of clocking in and

Table 3: Mining topic of Weibo in user education stage: distribution of feature words (excerpt).

Topic 1	Topic 2	Topic 3	Topic 4
Convenient	Clock in	One kilometer	Password
0.10314	0.03843	0.00942	0.00133
Give a like	Encounter	Green	Crack
0.08135	0.02414	0.00828	0.00128
Beautiful	Incredibly	Travel	Free
0.04432	0.01963	0.00764	0.00105
Like	Registered	Profit model	Hidden trouble
0.04012	0.01852	0.00732	0.00093
Big love	Expectation	Order	Deficit
0.34861	0.01284	0.00531	0.00081
Awesome	Ofo	Effectiveness	Locked
0.30234	0.00987	0.00397	0.00080
Fashion	Mobike	Health	Remove the seat
0.2396	0.00896	0.00381	0.00075
Cost-effective	App	Invention	Car chain
0.21753	0.00481	0.00362	0.00063

registering for the first time to use shared bikes. We call such topics public expectations. The Weibo with the topic tricluster is mostly the comments of media or opinion leaders on shared bikes. Such Weibo has a large number of words and is organized. We call this topic positive comments. Topic 4 of Weibo clustering contains a large number of characteristic words related to unlocking, password, and car lock, and the overall sentiment is relatively negative. This is because of cost control, technical constraints, and other factors. The mechanical locks used in the first-generation shared bicycles are easily cracked by users. At this time, many Weibo posts released the content of cracking the mechanical locks and riding for free. We classify such topics as having problems.

Topics 1, 2, and 3 show people's positive emotions towards shared bikes from different perspectives, while topic 4 shows people's negative emotions towards shared bikes. As shown in Figure 7, the proportion of microblogs with a different topic in the overall microblogs is as high as 92% with a positive attitude, which shows the public's love for shared bikes and tolerance of its shortcomings in this period.

5.2.2. Keyword Extraction and Analysis of Policy Text. During this period, there were no policies for bike sharing. The only two policy documents, Guiding Opinions on Promoting Green Consumption and National Fitness Plan (2016–2020), only included policies related to bike sharing to encourage cycling and low-carbon travel. The results show that words such as "the whole people," "fitness," "green," and "consumption" appear frequently. By looking at the sentences of these words, they are all policies guiding the whole industry. Therefore, it can be judged that in the early stage of bike-sharing development, there was no corresponding industrial policy support.

5.3. Demand Outbreak Stage Analysis

5.3.1. Weibo Topic Mining and Analysis. During this period, the number of microblogs increased sharply, including 31,259 Weibo. The topic-feature word distribution results of topic clustering are shown in Table 4. All microblogs are clustered into three different topics. Compared with the previous stage, the number of topics in the cluster decreased by one. This is because with the rapid expansion of bike-sharing enterprises, most cities have been covered, so the number of microblogs expecting the registration of bike-sharing enterprises decreased greatly. To be specific, the first topic is to express the convenience of daily use of shared bikes. The number of words on Weibo on this topic is relatively small, usually within a dozen words. Topic 2 is still the long microblog of media microblog number and opinion leaders' analysis and evaluation of shared bikes. It can be seen from the table that the wording of these Weibo is a formal and overall affirmation of shared bikes. Topic 3 is similar to topic 1, except that this topic expresses the problems encountered in daily use and makes fun of them.

The first topic of daily use and the second topic of positive comments can still be regarded as people's positive emotions towards shared bikes, while the third topic shows

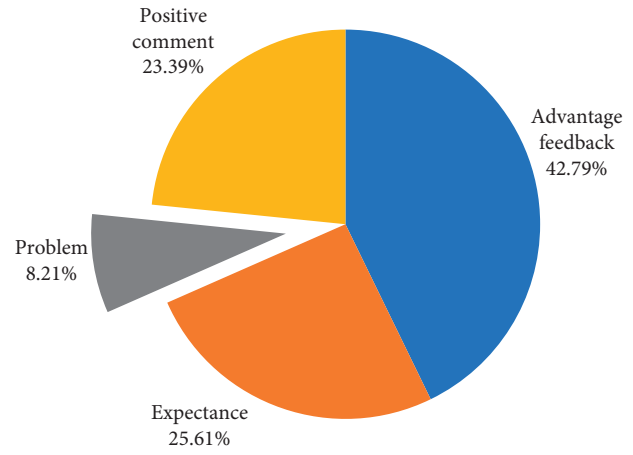

Figure 7: The proportion of the number of Weibo posts with different topics in the user training stage.

Table 4: Mining topic of Weibo in demand outbreak stage: distribution of feature words (excerpt).

Topic 1	Topic 2	Topic 3
Convenient	Short distance	QR code
0.15323	0.02071	0.10877
Free	Invention	Random parking
0.10334	0.01534	0.10060
Metro station	Pay	Locked privately
0.08641	0.01142	0.08265
Easily	Positioning	Deposit
0.06678	0.01087	0.06731
Reservation	Travel	Safe
0.06012	0.00881	0.05123
Praise	Lane	Damage
0.05621	0.00706	0.03876
Recommend	IPO	Broken
0.04130	0.00654	0.01134
Clock in	Quality	Phishing enforcement
0.02045	0.00586	0.00818

people's negative emotions towards shared bikes. As shown in Figure 8, compared with the previous stage, due to the sharp increase in the number of users and the number of bicycles, the corresponding service quality of enterprises, and the lack of government supervision in this period, the proportion of the problematic topics rose from 8.21% to 31.3%.

5.3.2. Keyword Extraction and Analysis of Policy Text. Policy documents directly related to shared bicycles have already appeared at this stage. After removing "Internet bicycles," "shared bicycles," "relevant departments," "operating companies," "vehicles," "transportation," and other words with high frequency but no practical meaning, the final keywords can be divided into two parts: one is the description of policy planning, such as "travel," "construction," and "standard," and the other is the description of specific policies, such as "parking" and "information". It also includes specific policy descriptions of "parking" and "information" related to practical issues of shared bikes.

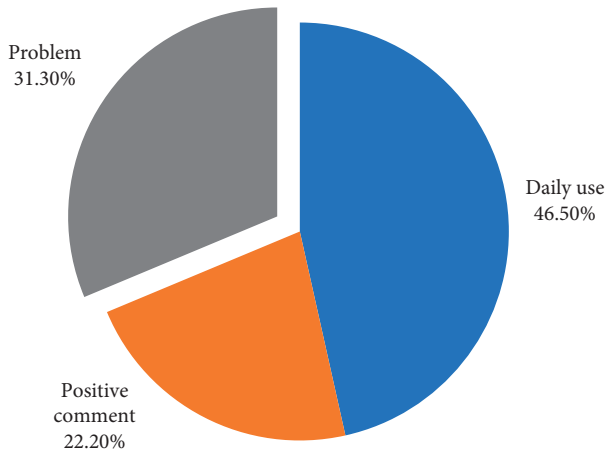

Figure 8: The proportion of the number of Weibo posts with different topics in the demand outbreak stage.

5.3.3. Comparison between Weibo Topics and Policy Texts. First, we compare the collinearity between the topic words and the high-frequency keywords. The top five topic keywords in the three probability of the topic are "QR code," "parking," "locking," "deposit," and "safety." The Weibo corresponding to these words concentrate on "QR code modification," "parking and placing," "locking without permission," "deposit trouble," and "driving safety." Policy highest frequency in the text of the five words: "travel," "park," "construction," and the "norm" and "information." These words reflect the government in policy-making for travel "convenience," "disorderly parking place," "service construction," "enterprise and individual behavior norms," and "personal user information security" issues such as the focus. Among them, some problems in Weibo topics and policy text exist at the same time, such as the simultaneous occurrence of keywords: "stop the place," and "locked" and "deposit" in the Weibo topic and issues such as policy text of "enterprise and individual behavior norms," "security" and "supporting service construction" proposed by the corresponding lanes. Although "two-dimensional code" can also be classified as the category of "personal code of conduct," but through the analysis of the original microblog, the problem of "two-dimensional code" described in the microblog contains the altered "two-dimensional code" fraudulent behavior. By comparison, the overall normative content of the policy is consistent with the problems excavated by Weibo, which indicates that the content of the policy formulation is in line with the basic demands of the public. Meanwhile, the absence of the problems represented by the high-probability feature word "QR code" in the policy text also reflects the lack of comprehensiveness of the policy text to some extent.

Second, we compared the relationship between the topic of the Weibo issue and the timeline of policy introduction. According to the statistics of the number of microblogs in the third topic every month, Figure 9 shows the change of the number of microblogs in the third topic in the eight months at this stage. It can be seen that the number of microblogs in January 2017 surged, which increased by more than four times compared with that in December 2016, and continued to increase in the following months. The first local bike-sharing encouragement and regulation policy was issued in Chengdu in March 2017, and the first national draft for soliciting opinions was issued in May. The time point for various regions to introduce policies on a large scale has already come from August to October. For the same problem, the time for the government to issue policies has a certain lag compared with the time for Weibo topics. Taking the national policy issued by the Ministry of Transport as the time node, the lag period is 4 months, and for the time node of the local policy issued by Chengdu, the lag period is 2 months. Obviously, for the emerging "Internet+" industry of bike sharing, which takes no more than 18 months to sprout, develop, and stabilize, a period of 2–4 months will leave a regulatory gap for the development of the industry, which will not only affect the user experience but also affect the development of enterprises and even the whole industry.

5.4. Elimination Stage Analysis

5.4.1. Weibo Topic Mining and Analysis. At this stage, the number of microblogs tended to stabilize and began to decline slowly, with a total of 27,627 microblogs. The topic-feature word distribution results of the Weibo topic clustering are shown in Table 5. 31,259 Weibo posts are clustered into five different topics. Compared with the previous stage, the number of topics in the clustering increased by two. The first topic is still to express the convenience of daily use of shared bikes, but its proportion is significantly reduced. This indicates that after a period of use, shared bikes have become a daily transportation tool for users, and users' interest in Weibo has decreased. Both topic 2 and topic 3 are comments on shared bikes. The difference is that the sentiment of topic 2 is optimistic, while the microblogs of topic 3 are mostly objective and neutral comments. The fourth topic is the problems encountered by users in daily use and ridicule. Most of the microblogs in topic 5 focus on the description of the deposit problem of shared bikes.

As shown in Figure 10, compared with the previous stage, the proportion of negative emotions is still on the rise. On the one hand, this is due to the decrease of daily clocked microblogs; on the other hand, the number of policies in the previous stage is small and the specificity is not enough, and the effect of regulation is limited. Notably, a new topic (topic 5) highlights the issue of deposits. Through checking the original microblog corresponding to topic 5, we find that the deposit problem mainly focuses on two aspects: one is that users need to pay multiple deposits when using different brands of shared bikes; the other is that some bike companies cannot withdraw deposits on time due to their blind expansion and lack of funds. Therefore, both topic 4 and topic 5 can be regarded as the existing problem topics of Weibo mining.

5.4.2. Keyword Extraction and Analysis of Policy Text. The proof policy of this stage country level already came on stage. From August to October, major cities rolled out local management policies for bike-sharing. After keyword extraction is removed, the high-frequency keywords include "positioning," "special account," "technology," and other words, indicating that the policies at this stage are more specific than those at the previous stage.

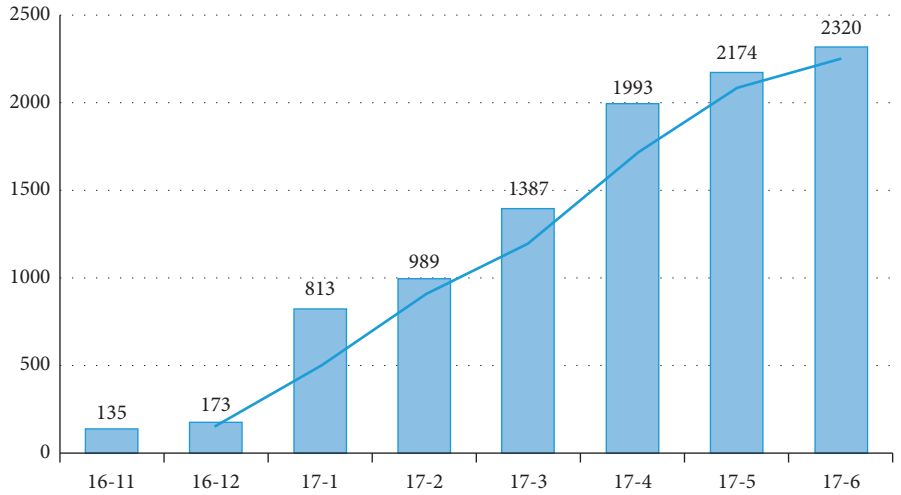

Figure 9: Monthly statistics on the number of Weibo posts of the third topic in the demand outbreak stage.

Table 5: Mining topic of Weibo in elimination stage: distribution of feature words (excerpt).

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Convenient 0.05452	Green 0.02071	System 0.00531	Random parking 0.01289	Deposit 0.06432
Free 0.04613	Travel 0.01534	Supervision 0.00519	Damage 0.1065	Zhima integral 0.0522
Applet 0.04171	Invention 0.01142	Regulatory 0.00408	Electric fence 0.0973	Account 0.04231
Easily 0.03983	Severe 0.01087	Protect 0.00391	QR code 0.0834	Extract 0.0387
Work 0.03286	Export 0.00881	12 years old 0.00353	Manned 0.0644	Rights 0.02231
Habit 0.02876	Information 0.00706	Real name 0.00298	Abandoned 0.0570	Credibility 0.02108
Campus 0.01971	Domestic 0.00654	Deficit 0.00223	Recycle 0.0445	Capital 0.01976
Shuttle 0.00886	Scenery 0.00586	Matching 0.00187	Disclosure 0.0386	Escape 0.00821

5.4.3. Comparison between Weibo Topics and Policy Texts. First, we compare the collinearity between the topic words and the high-frequency keywords. The top five topic words for topic 4 and topic 5 are "parking," "damage," "electronic fence," "deposit," and "Zhima points." The top five most frequently used words in policy texts were "positioning," "self-regulation," "special account," "illegal parking," and "technology." By comparison, it is found that "parking in disorder" and "illegal parking" are the same, "electronic fence" and "positioning" is basically the same, and "deposit" and "ant integral" and "special account" are all about the description of user deposit. Therefore, the content of the policy at this stage is consistent with the theme of microblog mining, which indicates that the content of the policy meets the basic demands of the public. To sum up, the above empirical evidence shows that the public online opinions represented by Weibo posts can have a certain influence on industrial policy formulation.

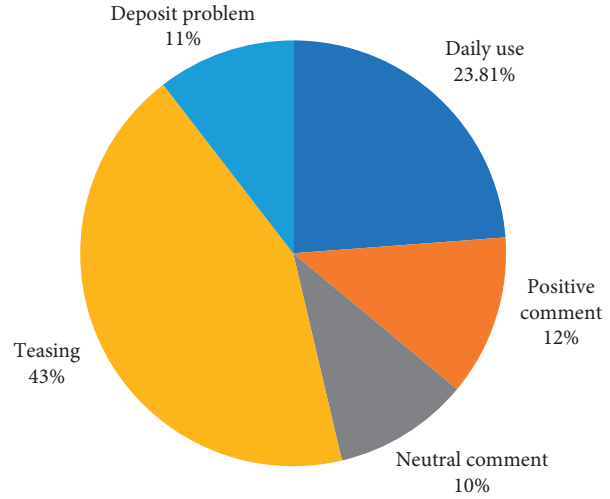

Figure 10: The proportion of the number of Weibo posts with different topics in the elimination stage.

Second, through the comparison of the timeline, we find that there is still a deviation between the time of the introduction of industrial policy and the peak time of network public opinion at this stage, with an average lag of 2 months. However, compared with the previous stage, this time has been reduced by half, which is related to the growth of the whole industrial chain and the tracking reports of other media (print media, TV media, and so on). The information released by users on Weibo has been spread and developed and has been concerned by the whole society.

6. Conclusion

First, a feature-enhanced short text clustering method is proposed to perform topic clustering on publicly derived big data. Second, keyword extraction based on word frequency is used to quantify the text of industrial policy. Finally, time is taken as the main line to analyze the co-occurrence of clustering topics and keywords. Through three different

stages of Weibo data mining and keyword analysis of policy text, we find the following.

In the stage of user training, bike sharing, as a typical "Internet+" industry, developed and grew in a short period by accurately solving the "pain points" of users and the power of capital. At this stage, more than 90% of microblogs expressed positive emotions for bike sharing, but at this stage, there was no effective industrial policy support.

In the stage of demand outbreak, with the increase of users, the fierce competition of manufacturers, the loss of bicycles, and the lack of supervision, the negative sentiment on Weibo at this stage accounted for 31.3%, which was nearly four times higher than the negative sentiment on Weibo at the previous stage, which was only 8%. Through the analysis of the timeline, we found that the time when the policy was issued had a certain lag compared with the problem time of public derivative big data mining represented by Weibo, and the regulatory vacuum caused by this lag led to the accumulation of negative emotions, which further affected the healthy development of the whole shared bike. Through the co-occurrence analysis of topic keywords and policy keywords, it is found that although the policy can cover most of the existing problems, it still omits hot issues like "QR code" in Weibo.

Finally, through the research of this paper, we find that the real situation of an industry can be reflected by collecting and clustering the information of industry on Weibo, and these reactions are highly correlated with the industrial policies of the industry. At the same time, we find that the formulation time of industrial policy is lagging behind that of the outbreak time of microblog information. For example, when it comes to industrial policy, proper consideration of the big data of network media represented by Weibo post will be conducive to the rapid introduction of industrial policy.

Data Availability

The data in this paper are all from the open information on the Internet, including two parts: the first part is the microblog data, which can be obtained from the website http://www.weibo.com; the second part is industrial policy documents from different government websites.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This study was supported by the Science and Technology Research Project of Jiangxi Provincial Department of Education (grant no. GJJ200318) and the International Social Science Fund General Project (grant no. 16BJY082).

References

[1] N. Zhang, "Analyzing public generated big data and restructuring government decision making process: review and prospect," *Chinese Public Administration*, vol. 10, pp. 19–24, 2015.

[2] B. Yu and K. Yang, "The interactive policy agenda-setting model in China's Internet incidents: an empirical study on events concerning social justice," *Journal of Nanjing Normal University (Social Science Edition)*, vol. 5, pp. 13–20, 2013.

[3] Z. Deng and Q. Meng, "We-media agenda setting: a new path to formation of public policy," *Journal of Public Management*, vol. 13, no. 2, pp. 14–22, 2016.

[4] H. Chen, R. H. L. Chiang, and V. C. Storey, *Business Intelligence and Analytics: From Big Data to Big Impact*, Society for Information Management and The Management Information Systems Research Center, South Minneapolis, MN, USA, 2012.

[5] N. Zhang, "Big data analysis of public derivatives and reconstruction of government decision-making process: theoretical evolution and research prospects," *Chinese Public Administration*, vol. 10, pp. 19–24, 2015.

[6] E. Johnston and Y. Kim, "Introduction to the special issue on policy informatics," *The Innovation Journal: The Public Sector Innovation Journal*, vol. 16, no. 1, pp. 1–4, 2011.

[7] E. G. Martin, R. H. MacDonald, L. C. Smith et al., "Policy modeling to support administrative decisionmaking on the New York State HIV testing law," *Journal of Policy Analysis and Management*, vol. 34, no. 2, pp. 403–423, 2015.

[8] M. I. Sirer, S. Maroulis, R. Guimerà, U. Wilensky, and L. A. N. Amaral, "The currents beneath the "rising tide" of school choice: an analysis of student enrollment flows in the Chicago public schools," *Journal of Policy Analysis and Management*, vol. 34, no. 2, pp. 358–377, 2015.

[9] K. A. Frank, W. R. Penuel, and A. Krause., "What is a "good" social network for policy implementation? The Flow of know-how for organizational change," *Journal of Policy Analysis and Management*, vol. 34, no. 2, pp. 378–402, 2015.

[10] L. Cai and X. Z. Yang, "Research on the application of big data in social public opinion monitoring and decision-making," *Administrative Tribune*, vol. 11, p. 75, 2021.

[11] H. Song, D. Peng, X. Huang, and J. Feng, "Research on Weibo hotspot finding based on self-adaptive incremental clustering," *Journal of Shanghai Jiaotong University*, vol. 24, no. 3, pp. 364–371, 2019.

[12] B. J. Ma, N. Zhang, and Q. Tan, "Analysis of influencing factors of public service effectiveness based on big data of interaction between government and people," *Chinese Public Administration*, vol. 400, no. 10, pp. 111–117, 2018.

[13] Y. Chun, Y. Kim, and H. Campbell, "Using Bayesian methods to control for spatial autocorrelation in environmental justice research: an illustration using toxics release inventory data for a Sunbelt county," *Journal of Urban Affairs*, vol. 34, no. 4, pp. 419–439, 2012.

[14] J. Li, Y. He, and Q. Xiong, "Research on network public opinion text mining based on big data technology," *Journal of Intelligence*, vol. 13, no. 10, pp. 1–6, 2014.

[15] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and algorithms*, Kluwer Academic Publishers, New York, NY, USA, 2002.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[17] J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 261–270, ACM, New York, NY, USA, February 2010.

[18] C. Honey and S. C. Herring, "Beyond microblogging: conversation and collaboration via twitter system sciences," in

Proceedings of the HICSS'09. 42nd Hawaii International Conference on, pp. 1–10, IEEE, Big Island, HI, USA, January 2009.

[19] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1445–1456, ACM, New York, NY, USA, May 2013.

[20] Q. Tang, *Short Text Clustering Based on BTM*, Anhui University, Hefei, China, 2014.

[21] Y. Zhang, *Short Text Similarity Calculation Based on Feature Extension of BTM Topic Model*, Anhui University, Hefei, China, 2014.

[22] A. Scrapy, *Fast and Powerful Scraping and Web Crawling Framework*, Scrapy. org. Np, Ballincollig, Ireland, 2016.