Hindawi
Complexity
Volume 2020, Article ID 9789431, 13 pages
https://doi.org/10.1155/2020/9789431

Research Article

Research on Sentiment Tendency and Evolution of Public Opinions in Social Networks of Smart City

Yanni Liu,[1] Dongsheng Liu ,[2] and Yuwei Chen[3]

[1]*School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China*
[2]*School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*
[3]*Beijing Yunzhenxin Technology Co., Ltd., Hangzhou 310012, China*

Correspondence should be addressed to Dongsheng Liu; lds1118@zjgsu.edu.cn

Received 29 March 2020; Revised 27 April 2020; Accepted 5 May 2020; Published 4 June 2020

Guest Editor: Zhihan Lv

With the rapid development of mobile Internet, the social network has become an important platform for users to receive, release, and disseminate information. In order to get more valuable information and implement effective supervision on public opinions, it is necessary to study the public opinions, sentiment tendency, and the evolution of the hot events in social networks of a smart city. In view of social networks' characteristics such as short text, rich topics, diverse sentiments, and timeliness, this paper conducts text modeling with words co-occurrence based on the topic model. Besides, the sentiment computing and the time factor are incorporated to construct the dynamic topic-sentiment mixture model (TSTS). Then, four hot events were randomly selected from the microblog as datasets to evaluate the TSTS model in terms of topic feature extraction, sentiment analysis, and time change. The results show that the TSTS model is better than the traditional models in topic extraction and sentiment analysis. Meanwhile, by fitting the time curve of hot events, the change rules of comments in the social network is obtained.

1. Introduction

With the wide application of the Internet technology, the Internet has gradually transformed to the dynamic platform for information sharing and interactive communication. The 43rd statistical report indicated that China had 854 million Internet users, and 99.1 percent of them access the Internet via mobile phones [1]. Social networks of a smart city have become the mainstream platform for information exchange and opinion expression. The users are not only the receivers of information, but also the creators to publish text comments in social networks. The hot events of public opinion refer that personal opinions are released on upcoming or already happened events by online communication tools and network platforms [2]. The spread of public opinion will snowball and expand by social networks, and emergent events may develop in an uncontrollable direction. Chain events caused by inadequate supervision on social networks can bring about the bad influence, and the frequency and harmfulness have shown an obvious rising trend in recent years [3].

Previous research has been studied from the qualitative aspects, such as the evolution mechanism of public opinion, information element classification, and influence judgment. However, the above research cannot meet the needs of online public opinion supervision, and the monitoring and management of hot events in social networks of a smart city need to implement quantitative judgment. For public opinion monitoring and management, Steyvers and Griffiths [4] proposed a topic model for public opinion detection in the social network. Yeh et al. [5] proposed a conceptually dynamic latent Dirichlet allocation (CD-LDA) model for topic content detection and tracking. The studies on probabilistic topic models for extracting hot topics from long texts have achieved good results [6], but these models are not suited to extract hot topics from short texts, such as Twitter and Facebook [7]. Kim et al. [8] introduced the sentiment scoring based on topics through the n-gram LDA

topic modeling technology and investigated the topic reports and sentiment dynamics in news about Ebola virus. Subeno et al. [9] proposed a collapsed Gibbs sampling method based on the latent Dirichlet allocation (LDA) model widely used on Spark. Park et al. [10] used partially collapsed Gibbs sampling for latent Dirichlet allocation and proposed a reasoning LDA method, which effectively obtained unbiased estimation under flexible modeling of heterogeneous text corpus by partially collapsed and Dirichlet mixed processing.

However, there are still some problems in the detection of public opinion for events in social networks of a smart city. Firstly, the detection and analysis of public opinions for hot events in social networks mostly remain in the qualitative analysis or empirical research, lacking quantitative research. Secondly, there is a lack of public opinion analysis method combining with the characteristics of the microblog in social networks. Thirdly, for the sentiment analysis of public opinion events, most research adopts the two-stage method. That is to detect the event firstly and then conduct sentiment analysis and judgment, which is likely to lead to the separation between the event and sentiment. Fourthly, the dissemination for public opinions of hot events is time sensitive, so it is necessary to involve the time factor in comment text analysis. Thus, this paper proposes a mixed model with dynamic topic-sentiment (TSTS) for short texts in the social network, which comprehensively incorporates the topics, sentiment, and time factor of events to detect the public opinion. By quantitative analysis of real experiment data, the model can not only show the quantitative evolution trend of public opinion, but also provide the propagation rule of sudden events.

The main contributions of this paper are reflected in two aspects. Firstly, the TSTS model is proposed by extending the topic model, which can not only extract both topic and sentiment polarity words, but also integrate the time factor to realize dynamic analysis of short texts. Secondly, this paper studies the detection and evolution analysis of Internet public opinion by the dynamic topic-sentiment model. The real datasets are used to conduct experimental analysis of the proposed model, which can reflect the evolution trend of public opinion diffusion.

2. Related Research

2.1. Research on Relevant Topic Models. The traditional opinion mining is to analyze the sentiment orientations based on the level of document and sentence. The traditional topic model was mainly used to compare the similarity among articles by comparing the number of repeated words in different articles. Blei et al. [11] proposed the latent Dirichlet allocation (LDA) topic model to mine the hidden semantics of the text. LDA is a three-layer Bayesian model involving the document, topic, and word. The document is composed of a mixed distribution of topics, and each topic follows a polynomial distribution. And, Dirichlet distribution is introduced as the prior information of the polynomial distribution. The schematic diagram of the LDA topic model is shown in Figure 1.

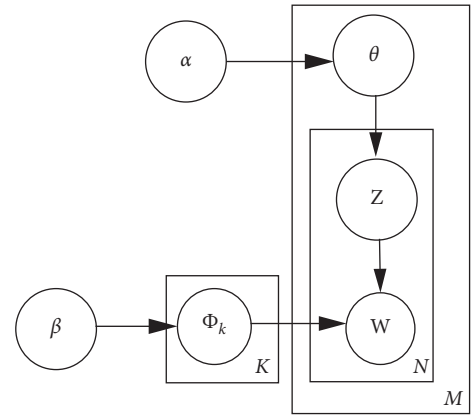

Figure 1: Schematic diagram of the LDA model.

2.1.1. Research on Topic Model of Short Text. For most topic models, the topic is the word that appears in the document and has some connection. In previous studies, the topic model of short texts was expanded by introducing relevant background or author information, which weakened the topic and produced meaningless word contributions. Similarly, if co-occurrence words are extended to the whole corpus in the experiment, the occurrence frequency of each word will be greatly increased, and the connection between words will be closer. Then, the modeling on documents will be easier. Based on the above hypotheses, Cheng et al. [12] proposed the biterm topic model (BTM), which is another way to explain the relationship between words, and text modeling of documents can be conducted based on the word co-occurrence model of the whole corpus. Rashid et al. [13] proposed the fuzzy topic modeling (FTM) for short texts from the fuzzy perspective to solve the sparsity problem. Based on BTM, Lu et al. [14] introduced the RIBS-Bigrams model by learning the usage relationship, which showed topics with two-letter groups. Zhu et al. [15] proposed a joint model based on the latent Dirichlet allocation (LDA) and BTM, which not only alleviates the sparsity of the BTM algorithm in processing short texts, but also preserves topic information of the document through the extended LDA.

2.1.2. Research on Mixed Model Integrating Topic Sentiment. To evaluate the sentiment tendency of documents, the joint sentiment topic (JST) model added the sentiment layer based on the LDA model, forming a four-layer Bayesian network [16]. In this structure, the sentiment polarity label is related to the document, and the word generation is also influenced by both topic and sentiment. In the traditional LDA model, the generation of the document and words is determined by the topic. But in the JST model, the word of the document is determined by the topic and the sentiment. Amplayo et al. [17] proposed the aspect and sentiment unification model (ASUM) with sentiment level. The difference between JST and ASUM is that words in a sentence come from different topics in the JST model, while all words of a sentence belong to one topic in the ASUM model.

2.1.3. Research on Topic Model with Time Factor. Yao et al. [18] revealed the semantic change process of words by correlating the time factor with Wikipedia text knowledge. In terms of event evolution, the associative topic model (ATM) is proposed [19], and the recognized cluster is represented as the word distribution of the cluster with the corresponding event. In addition, Topic Over Time (TOT) was proposed to integrate the time factor into the LDA model [20]. In the TOT model, word co-occurrence can affect the discovery of subject words, and time information can also affect the extraction of topic words. Unlike other models, each topic is subject to a continuous distribution of time and not rely on Markov models to discretize time in the TOT model. For each document generation, the mixed distribution of topics is determined by word co-occurrence and timestamp [21], which allows the TOT model to maintain independence in the time dimension and can predict the time for the document without any time information.

2.2. Gibbs Sampling. The derivation of the experimental model in the paper is a variant form of the Markov Chain, so the Markov Chain Monte Carlo (MCMC) method is used for sampling in the experiment. Gibbs sampling, as one of the MCMC methods, has been widely used in prior research. Gibbs sampling is used to obtain a set of observations that approximate a specified multidimensional probability distribution, such as the probability distribution of two random variables.

The Gibbs sampling method used for the latent Dirichlet allocation (LDA) model can significantly improve the speed of the real-text corpus [22]. Papanikolaou et al. [23] estimated latent Dirichlet allocation (LDA) parameters from Gibbs sampling by using all conditional distributions of potential variable assignments to effectively average multiple samples. Zhou et al. [24] proposed two kinds of Gibbs sampling inference methods, such as Sparse BTM and ESparse BTM, to achieve BTM by weighing space and time. Bhuyan [25] proposed the correlation random effect model based on potential variables and an algorithm to estimate correlation parameters based on Gibbs sampling.

3. Model Constructing

3.1. Topic-Sentiment Mixture Model with Time Factor (TSTS). Based on prior research, this paper mainly improves the topic model from three aspects. Firstly, the sparse matrix caused by short texts in the social network is solved. Secondly, the topic and sentiment distribution of the same word pair are controlled. Thirdly, the problem of text homogeneity is solved by incorporating the time factor into the topic model. Therefore, the TSTS model proposed in this paper is used to constrain the word pairs in the same document, which greatly reduces the complexity of time and space and makes up for the sparse matrix of short texts to some extent. Moreover, the sentiment layer is integrated into TSTS by extending the hypotheses of ASUM and restrains word pairs generated by constraining sentences to follow the same topic-sentiment distribution. Finally, the TSTS model incorporating the time factor does not rely on the Markov model to discretize time, and each topic is subject to the continuous temporal distribution. For each document generation, the mixed distribution of topics is determined by the words co-occurrence and timestamps. TSTS model is shown in Figure 2.

The TSTS model simulates the generating process of online comments. Generally, the online comments from users can be regarded as a document, which is short, pithy, and highly emotional. The word co-occurrence from BTM is the most effective solution for the short-text topic model. In addition, the TSTS model with the time layer can continuously sample users' evaluation of hot events, as well as the dynamic changes of users' sentiment. Therefore, the hypotheses of the TSTS model are proposed as follows:

(i) The probability distribution of the time factor is not directly equal to the joint distribution of the topic and sentiment

(ii) The topic-sentiment distribution of each document is independent [26]

(iii) Similar topics of different sentiment polarity are not automatically categorized [27]

Combined with the probability graph of Bayesian's network, the TSTS model proposed in the paper has four characteristics. First, a word pair is used to replace a single word to carry out the sampling model. Second, each timestamp is related by topic and sentiment. Third, the extraction of thematic characteristic and sentiment words is for the whole corpus. Fourth, in the derivation process of the TSTS model, it is not necessary to correspond between thematic feature and affective polarity words. That is because every topic and sentiment have the corresponding polynomial word pair distribution. In addition, the text modeling process of the TSTS model also follows the assumption that there is a connection between the sentimental polarity words of the topic features, which also changes with the time factor. So, the documents used to train the model must have a specific timestamp, such as the publishing time of the microblog.

3.2. Generation of a Text in TSTS Model. In the TSTS model, we assumed that a corpus is composed of several texts. For instance, a microblog is a text containing two dimensions of topic and sentiment. Considering the effectiveness of public opinions and related parameters of the microblog text, word distribution is determined by the topic, sentiment, and time. So, TSTS is an unsupervised topic-sentiment mixed model. The generation process of the document is as follows:

(1) Extract a polynomial distribution θ_d on a topic from the Dirichlet prior distribution α, that is, $\theta_d \sim \mathrm{Dir}(\alpha)$

(2) Extract a polynomial distribution $\psi_{z,l}$ at some point from the Dirichlet prior distribution μ, that is, $\psi_{z,l} \sim \mathrm{Dir}(\mu)$

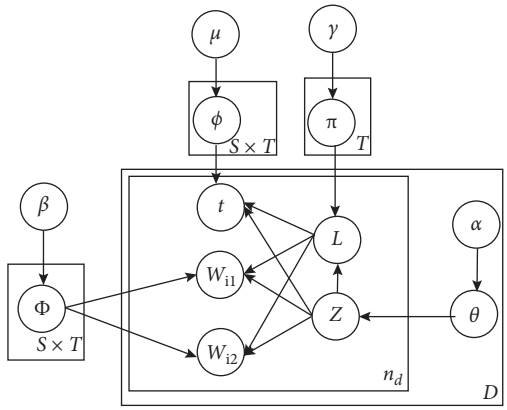

Figure 2: TSTS model.

(3) Extract a polynomial distribution π_z in a sentiment from the Dirichlet prior distribution γ, that is, $\pi_z \sim \mathrm{Dir}(\gamma)$

(4) For each document d and for each pair of words in the article $b = (w_{i1}, w_{i2}), b \in B$,

(a) Choose a topic $z_i \sim \theta_d z_i \sim \theta_d$
(b) Choose an emotional label $l_i \sim \pi_{z_i}$
(c) Choose a pair of words $b_i \sim \varphi_{z_i,l_i}$
(d) Choose a timestamp $t_i \sim \psi_{z_i,l_i}$

As shown in Figure 2, word pairs in a document may belong to different timestamps in the text generation process of the TSTS topic model. In theory, all the content of an article such as words and topics should belong to the same timestamp. Also, the introduction of the time factor into the topic model will affect the topic homogeneity of an article. However, the default time factor of the TSTS model in the topic model will not affect the homogeneity of the text. So, it is assumed that the time factor in the paper has no weight. Based on the TOT and the group topic (GT) model, the superparameter μ is introduced into TSTS to balance the interaction of time and words in document generation. The parameters' explanation of the TSTS model is shown in Table 1.

3.3. Model Deduction. According to the Bayesian network structure diagram of the TSTS model, the polynomial distribution $\boldsymbol{\theta}$ of the topic, the distribution $\boldsymbol{\pi}$ of sentiment with the topic, the correlation distribution $\boldsymbol{\varphi}$ of word pairs with <topic, sentiment>, and the correlation distribution $\boldsymbol{\psi}$ of time with <topic, sentiment> can be calculated according to the superparameters α, β, γ, and μ. Then, Gibbs sampling is done that can ensure the convergence of the TSTS model under enough iteration times. And, each word in the document is assigned the topic and sentiment that are most suitable for the facts.

According to the principle of Bayesian independence, the joint probability of word pair, topic, sentimental polarity, and timestamp is given as follows:

$$p(\mathbf{b}, \mathbf{t}, \mathbf{l}, \mathbf{z} \mid \alpha, \beta, \gamma, \mu) = p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \beta) \cdot p(\mathbf{t} \mid \mathbf{l}, \mathbf{z}, \mu) \cdot p(\mathbf{l} \mid \mathbf{z}, \gamma) \cdot p(\mathbf{z} \mid \alpha), \quad (1)$$

Table 1: Explanation of parameters.

D	Number of documents
V	Vocabulary size
T	Number of topics
S	Number of sentiment polarity
H	Number of timestamps
M	Number of word pairs
B	Set of word pairs
B	Word pairs, $b = (w_{i1}, w_{i2})$
W	Word
T	Time
Z	Topic
L	Sentiment polarity label
Θ	$[\theta_d]$: polynomial distribution of topics
Φ	$[\varphi_{z,l}]$: $T \times S \times V$ matrix, word pairs' distribution
Π	$[\pi_z]$: $T \times S$ matrix, sentiment distribution
Ψ	$[\psi_{z,l}]$: $T \times S \times H$ matrix, time distribution
α	Dirichlet prior parameters of Θ
γ	Dirichlet prior parameters of π
β	Asymmetric Dirichlet prior parameters of Φ
μ	Dirichlet prior parameters of ψ
n_d	The number of word pairs in document d
$n_{d,j}$	The number of word pairs for topic j in document d
n_j	The number of word pairs for topic j
$n_{j,k}$	The number of word pairs assigned as topic j and sentiment polarity k
$n_{i,j,k}$	The number of word pair b_i is assigned to the topic j and sentiment polarity k
$n_{j,k,h}$	The number of word pair b_i is assigned to the topic j and sentiment polarity k when timestamp is h
n^{-p}	The number of word pairs in the current document except for the p position

where the parameters are independent such as word pairs $\mathbf{b}$ and parameters α, γ, and μ, timestamps $\mathbf{t}$ and parameters α, γ, and β, sentiment polarity $\mathbf{l}$ and parameters α, μ, and β, and topic words $\mathbf{z}$ and parameters β, γ, and μ. Therefore, the joint distribution in the equation can be obtained by calculating the four parts on the right side of the equation.

Given the sentiment polarity label of specific topic features, the distribution of $\mathbf{b}$ can be regarded as a polynomial distribution. Based on the premise of topic words z_i and l_i, b_i is generated by N times with the probability $p(\mathbf{b} \mid \mathbf{l}, \mathbf{z})$ at each time. Given that word pairs are independent of each other, we can obtain

$$p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \beta) = \prod_{i=1}^{N} p\left(b_i \mid z_i, l_i\right) = \prod_{i}^{N} \beta \cdot b_i. \quad (2)$$

Superparameters are the representation parameters of the framework in the machine learning model [28], such as the number of classes in the clustering method or the number of topics in the topic model. In the Bayesian network, the distribution and density function of $\boldsymbol{\theta}$ are denoted as $H(\theta)$ and $h(\theta)$, respectively. They are regarded as the prior distribution function and the prior density function, respectively, which are collectively referred to as the prior distribution. If the distribution of $\boldsymbol{\theta}$ is obtained after sampling, it is called the posterior distribution. Based on the conjugate property of Dirichlet~multinomial, when the

parameters in the population distribution conform to the distribution law of polynomial (Multinomial), the conjugate prior distribution conforms to the following distribution:

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) + \mathrm{Mult}(\boldsymbol{\delta}) = \mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha} + \boldsymbol{\delta}). \tag{3}$$

For the general text model, the discretized Dirichlet distribution and multinational distribution are as follows:

$$\mathrm{Dir}(\mathbf{b} \mid \beta) = \frac{\Gamma\left(\sum_{j=1}^{T} \beta\right)}{\prod_{j=1}^{T} \Gamma(\beta)} \prod_{j=1}^{T} n_j, \tag{4}$$

$$\mathrm{Mult}(n \mid \mathbf{b}, N) = \binom{N}{n} \prod_{j=1}^{T} n_j, \tag{5}$$

where i, j, k, and h represent the iteration times of word pairs, topic, sentiment, and timestamp in the modeling process, respectively. Since the distribution of $p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \beta)$ follows the Dirichlet distribution, this paper introduces φ for $p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \beta)$. It can be obtained by integrating φ:

$$\begin{aligned} p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \beta) &= \int p(\mathbf{b} \mid \mathbf{l}, \mathbf{z}, \varphi) \cdot p(\varphi \mid \beta) \mathrm{d}\varphi \\ &= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{T \cdot S} \prod_j \prod_k \frac{\prod_i \Gamma\left(n_{i,j,k} + \beta\right)}{\Gamma\left(n_{j,k} + V\beta\right)}. \end{aligned} \tag{6}$$

To estimate the posterior parameters φ in the formula, we can combine with the Bayes formula and the conjugate property of Dirichlet~multinomial. The distribution of the posterior parameters can be obtained as follows:

$$p((\varphi \mid \mathbf{l}, \mathbf{z}, \beta)) \propto \mathrm{Dir}\left(\varphi \mid n_{i,j,k} + \beta\right). \tag{7}$$

Given that the expectation of the Dirichlet distribution is $E(\mathrm{Dir}(\varepsilon)) = \varepsilon_i / \sum_i \varepsilon_i$, so the calculated parameters are estimated by the known posterior parameter distribution expectation. The estimated results are shown in equation (7). Similarly, for $p(\mathbf{t} \mid \mathbf{l}, \mathbf{z}, \mu)$, ψ is introduced. By integrating ψ, it can be obtained as follows:

$$p(\mathbf{t} \mid \mathbf{l}, \mathbf{z}, \mu) = \left(\frac{\Gamma(H\mu)}{\Gamma(\mu)^H}\right)^{T \cdot S} \prod_j \prod_k \frac{\prod_h \Gamma\left(n_{j,k,h} + \mu\right)}{\Gamma\left(n_{j,k} + H\mu\right)}. \tag{8}$$

For $p(\mathbf{l} \mid \mathbf{z}, \gamma)$, π is introduced. By integrating π, it can be obtained as follows:

$$p(\mathbf{l} \mid \mathbf{z}, \gamma) = \left(\frac{\Gamma\left(\sum_k \gamma_k\right)}{\prod_k \Gamma(\gamma_k)}\right)^{T} \prod_j \frac{\prod_k \Gamma\left(n_{j,k} + \gamma_k\right)}{\Gamma\left(n_j + \sum_k \alpha_k\right)}. \tag{9}$$

For $p(\mathbf{z} \mid \alpha)$, θ is introduced. By integrating θ, it can be obtained as follows:

$$p(\mathbf{z} \mid \alpha) = \left(\frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma\left(\alpha_j\right)}\right)^{D} \prod_d \frac{\prod_j \Gamma\left(n_{d,j} + \alpha_j\right)}{\Gamma\left(n_d + \sum_j \alpha_j\right)}. \tag{10}$$

The TSTS model can estimate the posterior distribution after estimated values z and s have been obtained by sampling calculations. Then, the calculated equations (2)–(6) are brought into equation (1). Combining with the nature of Gamma function, the conditional distribution probability in Gibbs sampling can be obtained:

$$\begin{aligned} &p\left(s_p = k, z_p = j \,\middle|\, \mathbf{b}, \mathbf{t}, \mathbf{l}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma, \mu\right) \\ &\propto \frac{n_{d,j}^{-p} + \alpha_j}{n_d^{-p} + \sum_j \alpha_j} \cdot \frac{n_{w_p,j,k}^{-p} + \beta}{n_{j,k}^{-p} + V\beta} \cdot \frac{n_{j,k}^{-p} + \gamma_k}{n_j^{-p} + \sum_k \gamma_k} \cdot \frac{n_{j,k,t_p}^{-p} + \mu}{n_{j,k}^{-p} + H\mu}. \end{aligned} \tag{11}$$

In order to simplify equation (6), the superparameter $\mu = 1/n_d$ is introduced. When the superparameters α, β, μ, and γ are given, the set B of the word pair, the corresponding topic z, and sentiment label l can be used to infer the parameters φ, θ, π, and ψ based on Bayes' rule and Dirichlet conjugate properties:

$$\begin{aligned} \varphi_{j,k,i} &= \frac{n_{i,j,k} + \beta}{n_{j,k} + V\beta}, \\ \theta_{d,j} &= \frac{n_{d,j} + \alpha_j}{n_d + \sum_j \alpha_j}, \\ \pi_{j,k} &= \frac{n_{j,k} + \gamma_k}{n_j + \sum_k \gamma_k}, \\ \psi_{j,k,h} &= \frac{n_{j,k,h} + \mu}{n_{j,k} + H\mu}. \end{aligned} \tag{12}$$

4. Experiment Analysis

4.1. Data Collection. In order to verify the TSTS model proposed in this paper, the four hot events are randomly selected from the trending searches of Sina Weibo in 2019. And, the comments of four events are regarded as the experimental datasets. The four datasets selected are "Military parade in National Day," "The assault on a doctor," "Hong Kong's event," and "Garbage sorting in Shanghai." The comments are extracted from the Sina social network platform. In the original datasets, there are some meaningless words in the microblog text, such as stop words, interjections in tone, punctuation marks, and numeric expressions. Before text modeling, the word segmentation package in Python is used to process the experimental initial dataset. In addition, considering that comments on social networks are relatively new, the fashionable expressions in the social network are collected and added to the customized dictionary. So, these emerging words can be identified as far as possible and replaced with normal expressions. In addition, there are some useless words in the text, such as URL links and numbers, which can be filtered by regular expressions. Finally, a total of 14288 experimental data in four events are obtained. The description of four datasets is shown in Table 2.

4.2. Sentiment Dictionary. The words or phrases in the sentiment dictionary have obvious sentiment tendency,

Table 2: Experiment datasets.

The dataset (the number of comments)	Number of words in per microblog		Vocabulary size	
	Initial	Pretreatment	Initial	Pretreatment
The dataset 1 (3562)	134	102	9789	6319
The dataset 2 (3527)	127	94	9736	6242
The dataset 3 (3617)	131	100	9780	6301
The dataset 4 (3582)	128	96	9742	6254
Average	130	98	9762	6279

which can be divided into positive and negative words. The sentiment dictionary in this paper has two major roles. On the one hand, we can identify sentiment polarity words and distinguish topic features and sentiment words. On the other hand, combining with sentiment prior information to make the model more accurate in judging the sentiment polarity of the text. Given that sentiment polarity words can reflect users' sentiment tendency, it is of great significance to analyze the sentiment orientation of the text.

At present, there are two major Chinese sentiment dictionaries: NTU and HowNet. The former dictionary contains 2812 positive words and 8276 negative words. The latter contains about 5,000 positive words and 5,000 negative words. Based on HowNet and the classification of sentiment polarity [29, 30], this paper constitutes the sentiment dictionary of the TSTS model evaluation experiment, as shown in Table 3.

4.3. Parameter Setting. In this paper, the Gibbs algorithm is used to sample the TSTS model and estimate four posterior parameters. According to the parameter setting in the traditional topic model, the superparameters are set as follows. First, the superparameter α is set as $50/K$, and K is the number of topics extracted. Second, β is set as 0.01. Third, γ is set as $(0.05 \times \mathrm{AVE})/S$. AVE stands for the average length of articles, that is, the average number of words in the microblog in this experiment, and S stands for the total number of polar tags. Finally, μ is set as $1/n_d$.

4.4. Evaluation Indicator. For the extraction of topic features, perplexity is used as an evaluation indicator to measure the predictive power of unknown data in the process of model modeling. Also, the lower perplexity means better efficiency. The calculation formula of the perplexity is as follows:

$$\text{perplexity} = P\left(\widetilde{D}_t \mid \mathscr{M}\right) = \exp\left\{-\frac{\sum_{d=1}^{D^t} \log P\left(\widetilde{b}_d^t \mid \mathscr{M}\right)}{\sum_{d=1}^{D^t} \widetilde{N}_d^t}\right\}, \tag{13}$$

where $\widetilde{D}_t = \left\{\widetilde{b}_d^t\right\}_{d=1}^{D^t}$ represents an unknown dataset with the timestamp t.

$$P\left(\widetilde{b}_d^t \mid \mathscr{M}\right) = \prod_{n=1}^{\widetilde{N}_d^t} \prod_{l=1}^{L} \prod_{t=1}^{T} P\left(\widetilde{b}_{d,n} \mid l, z\right) P(z \mid l) P(l), \tag{14}$$

Table 3: Classification of sentiment words.

Sentiment labels	Happy	Surprise	Sad	Angry
Vocabulary size	2467	276	3025	1897

where $\widetilde{b}_d^t$ represents the vector set of word pairs in text d, $\widetilde{N}_d^t$ represents the number of word pairs in $\widetilde{b}_d^t$, and $P(\widetilde{b}_d^t \mid \mathscr{M})$ represents the direct possibility of training corpus, and the formula is as follows:

$$P\left(\widetilde{b}_d^t \mid \mathscr{M}\right) = \prod_{i=1}^{V} \left(\sum_{l=1}^{l} \sum_{z=1}^{T} \varphi_{l,z,i} \cdot \theta_{d,l,z} \cdot \pi_{d,l}\right)^{\widetilde{N}_{d,i}^t}. \tag{15}$$

For sentiment segmentation, the sentiment judgment from the perspective of the document is used as the evaluation index, which is based on the sentim ent polarity label in the sentiment dictionary. For the documents in this experiment, the positive and negative sentiment of a document can be judged. This paper adopts the consistency test method to mark the sentiment labels [31].

5. Results

5.1. Extraction of Topics. The primary task of the TSTS model is to extract topic features. As an extension of the topic-sentiment mixed model, the assessment is to judge whether the extracted topic features are reasonable and accurate. Before extracting topic features from text modeling, it is necessary to determine the number of topics to be extracted and the iteration times of Gibbs sampling. For the effective evaluation of topic discovery, the degree of perplexity is used as the measurement index in the paper. The lower the perplexity is, the better the fitting effect of the model is. Taking dataset 1 as an example, the simulation results are shown in Figure 3.

Based on the experimental results shown in Figure 3, the number of topics was set 20 in the subsequent experiments. In addition, we can calculate the perplexity of three models with the change of the iterations. By comparing the experimental results of TSTS and LDA, it can be found that the effect of TSTS is always better than LDA, and the degree of perplexity decreases with the increase of iteration. That indicates that the topic discovery ability of TSTS gradually improves, mainly because the TSTS model incorporates the word pairs to alleviate the sparse matrix of LDA for short texts. By comparing the experimental results of TSTS and BTM, it can be found that TSTS was better than BTM when the number of iterations increases. However, as the number

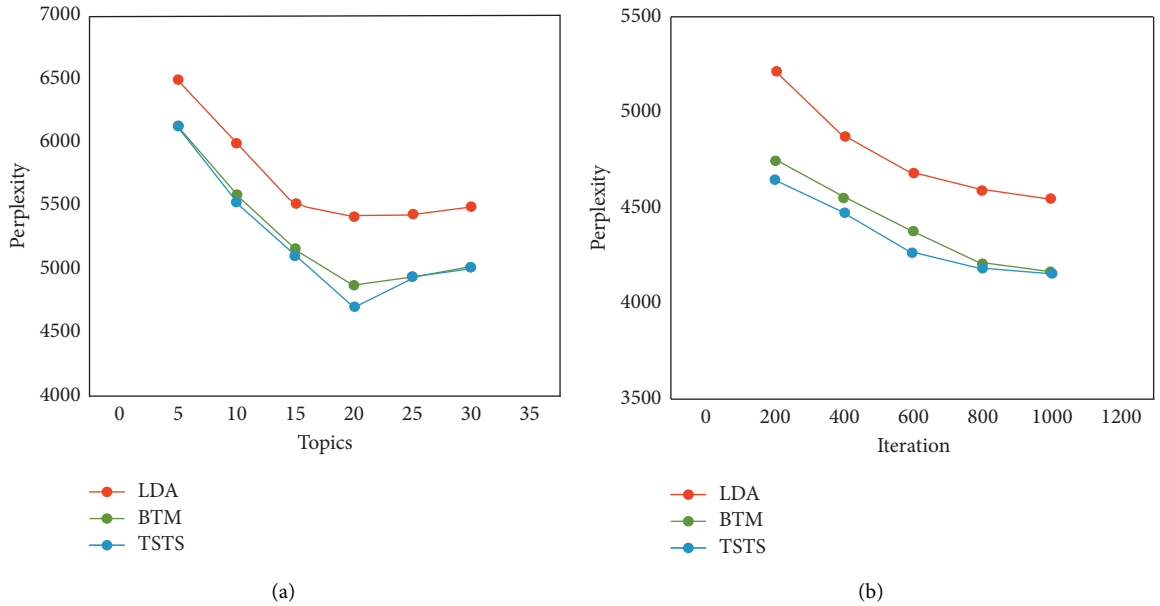

Figure 3: The relationship of the perplexity with the number of (a) topics and (b) iterations.

of iterations increased, the gap of both models became smaller. This is because the word pair of BTM is used for the whole corpus. When the number of iterations is small, the proportion of noise words is relatively large, resulting in poor quality of topic words. In addition, the sentiment layer is integrated into TSTS, and the error generated in the sentiment estimation will affect the next iteration. Although TSTS is worse than BTM when there are more iterations, the effect of TSTS can still be balanced with BTM. Therefore, during the extraction of topic features, the number of topics and iterations can be set as 20 and 600.

5.2. Sentiment Polarity. The information related to sentiment polarity is provided in accordance with the topic and sentiment polarity of words. The sentiment distribution of topics extracted from the TSTS model is shown in Figure 4. In addition, JST and ASUM are introduced as the comparison to measure the effect of sentiment recognition of the TSTS model. Each document has a binary sentiment label, such as positive or negative sentiment. Taking dataset 2 "The attack on the doctor" as an example, the result is shown in Figure 4. The number of topics is set to 5 at the beginning of the experiment. With the refinement of granularity, the performance of the TSTS model increases. Compared with JST and ASUM, the curve of the TSTS model changes greatly considering the topic and sentiment relationship among word pairs of the document. The change curve of the JST model shows a steady upward trend, and the identification efficiency of ASUM is low. That is because ASUM has strict assumptions, and the increase in the number of topics will cause the decentralization of topics and sentiments, which has a great negative impact on the overall performance of the model. The overall effect of the TSTS model was slightly better than JST and ASUM, but the effect decreased slightly after the number of topics increased to 20. This is because the

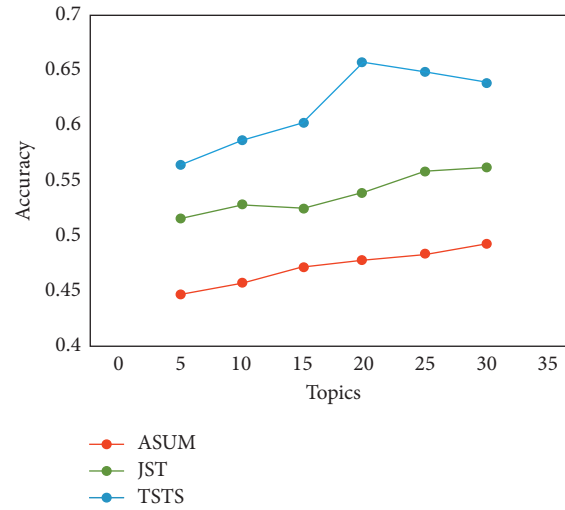

Figure 4: Accuracy of sentiment polarity judgment.

data collected in the dataset are limited, and the number of topics has been set to discretize the word distribution. Thus, the judgment of sentiment polarity is affected. The sentiment label classification of documents is compared under different topics, and the result of the TSTS model is better than JST and ASUM.

With the increase of topics, the recognition performance of the topic model has some fluctuations. But, the TSTS model was always better than JST and ASUM. When the number of obtained topics and iterations is set as 20 and 600, the TSTS is the best model in topic detection. When the number of topics in the four datasets was set as 20, the accuracy of sentiment polarity judgment is shown in Table 4.

From Table 4, the TSTS model is better than JST and ASUM in judging the sentiment polarity of documents. This

Table 4: Accuracy of sentiment polarity judgment.

	ASUM	JST	TSTS
The dataset 1	0.4763	0.5427	0.6348
The dataset 2	0.4617	0.5398	0.6599
The dataset 3	0.4832	0.5461	0.6475
The dataset 4	0.4841	0.5294	0.6522

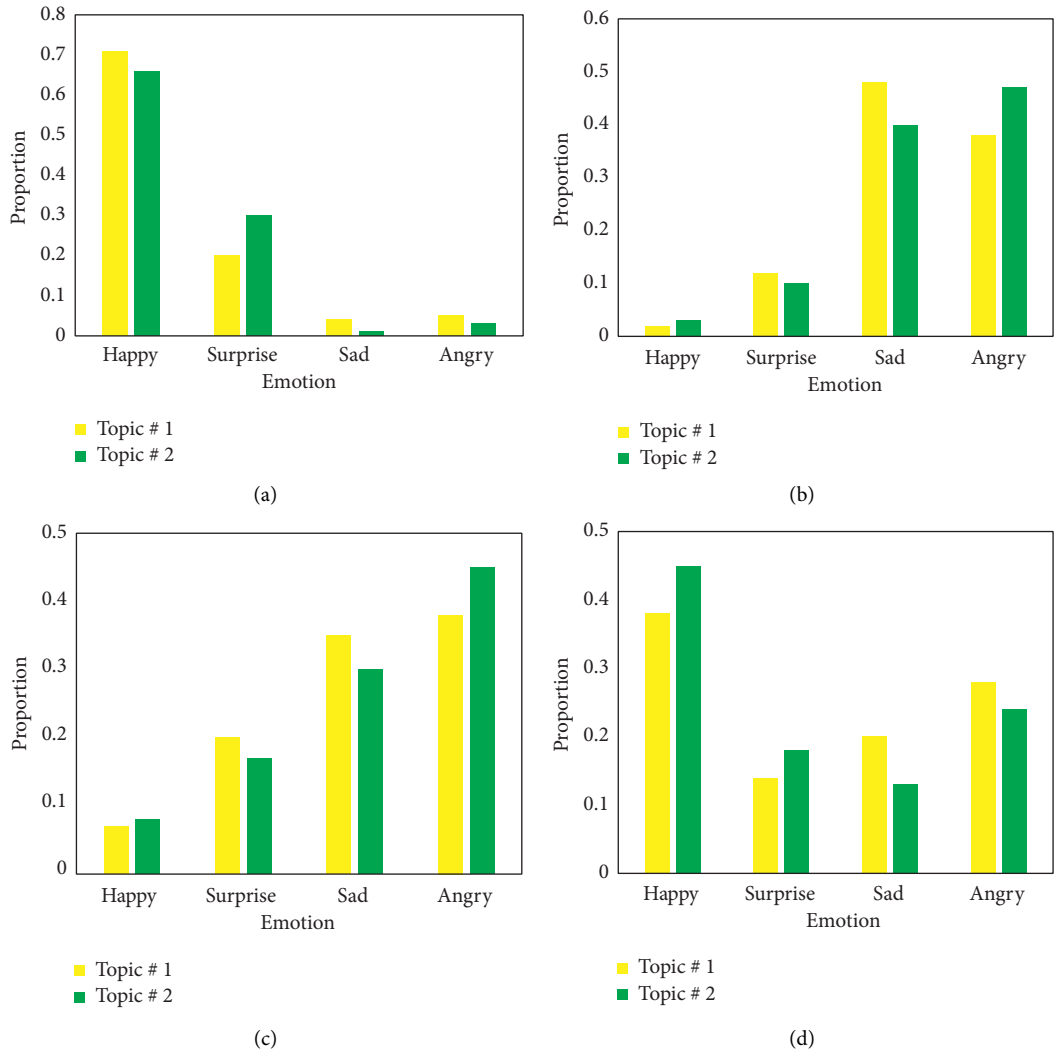

Figure 5: Sentiment distribution in four datasets. (a) Military parade in National Day. (b) The assault on a doctor. (c) Hong Kong's event. (d) Garbage sorting in Shanghai.

is because sentiment polarity depends on the performance of topic discovery in the previous stage. In this experiment, the effect of JST and ASUM was exactly opposite. The difference was caused by the length of the original document, which also indirectly verified the effectiveness of the TSTS model.

From Figure 5, it can be seen that the proportion of positive sentiment is significantly higher than the other sentiments in the dataset "Military parade in National Day" and the dataset "Garbage sorting in Shanghai," which is consistent with the sentiment tendency of users in the social network. For the second dataset "The assault on a doctor," the two kinds of negative sentiment polarities of topic #1 and topic #2 were compared. Topic #1 is more likely to be sad sentiment, while topic #2 is more likely to be angry sentiment. Topic #1 reflects the statement of the event, and topic #2 represents the follow-up discussion of the event.

5.3. Topic and Sentiment Evolution. The curves of topic features extracted from four datasets through the TSTS model are shown in Figure 6. Taking dataset 2 as an example, the topic curve conforms to the evolution law of social and abrupt events. The two curves represent the trend of feature words over time in topic #1 and topic #2. Topic #1 is the

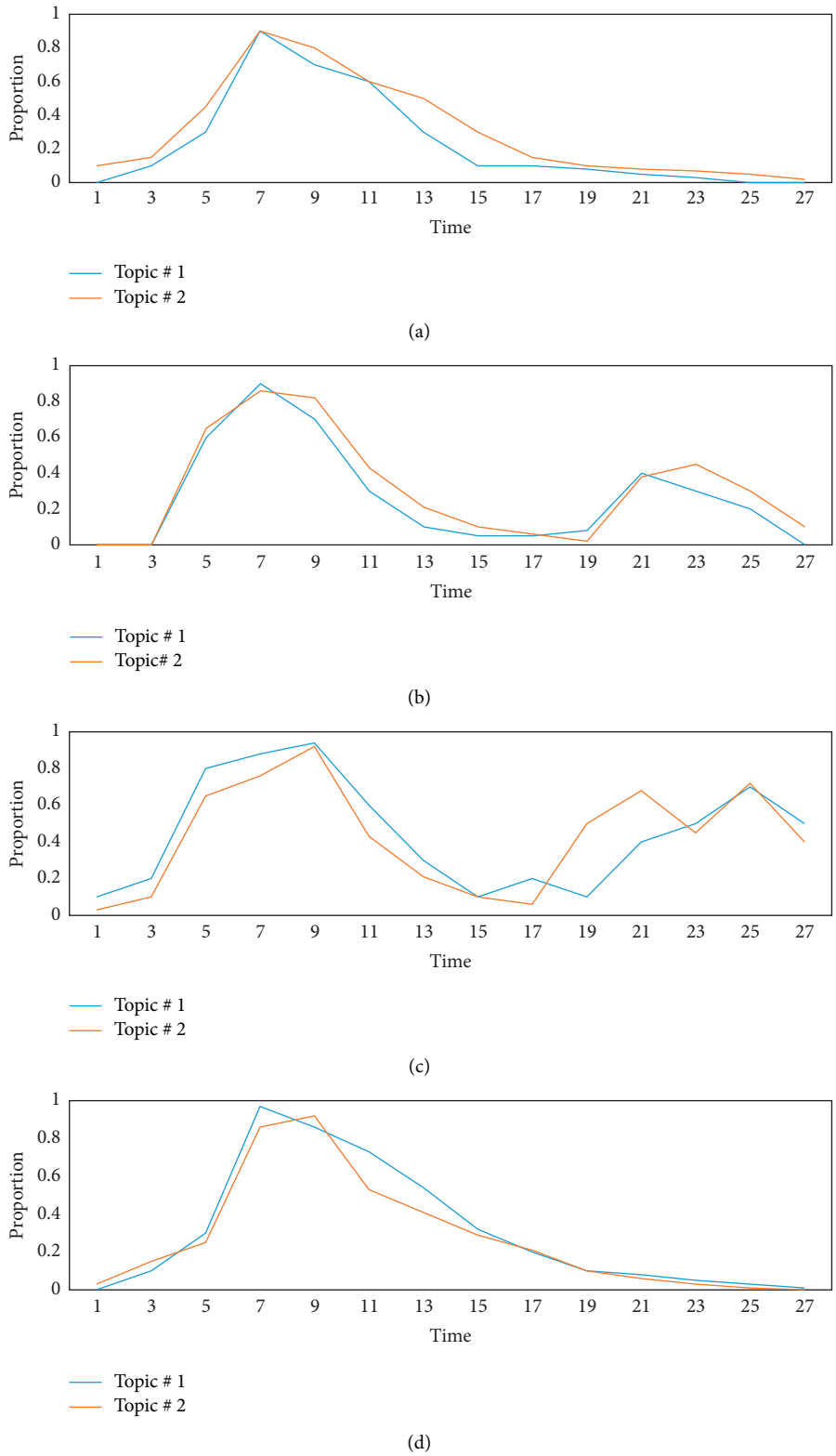

Figure 6: The changing of topics in four datasets. (a) Military parade in National Day. (b) The assault on a doctor. (c) Hong Kong's event. (d) Garbage sorting in Shanghai.

statement about the case itself. From the beginning of the event, the amount of discussion about the event on the social network rose sharply and then gradually declined. Topic #2 is a discussion on the development of the case, which has caused the second hot discussion again. The time is not consistent when two curves reach the high peak. The peak

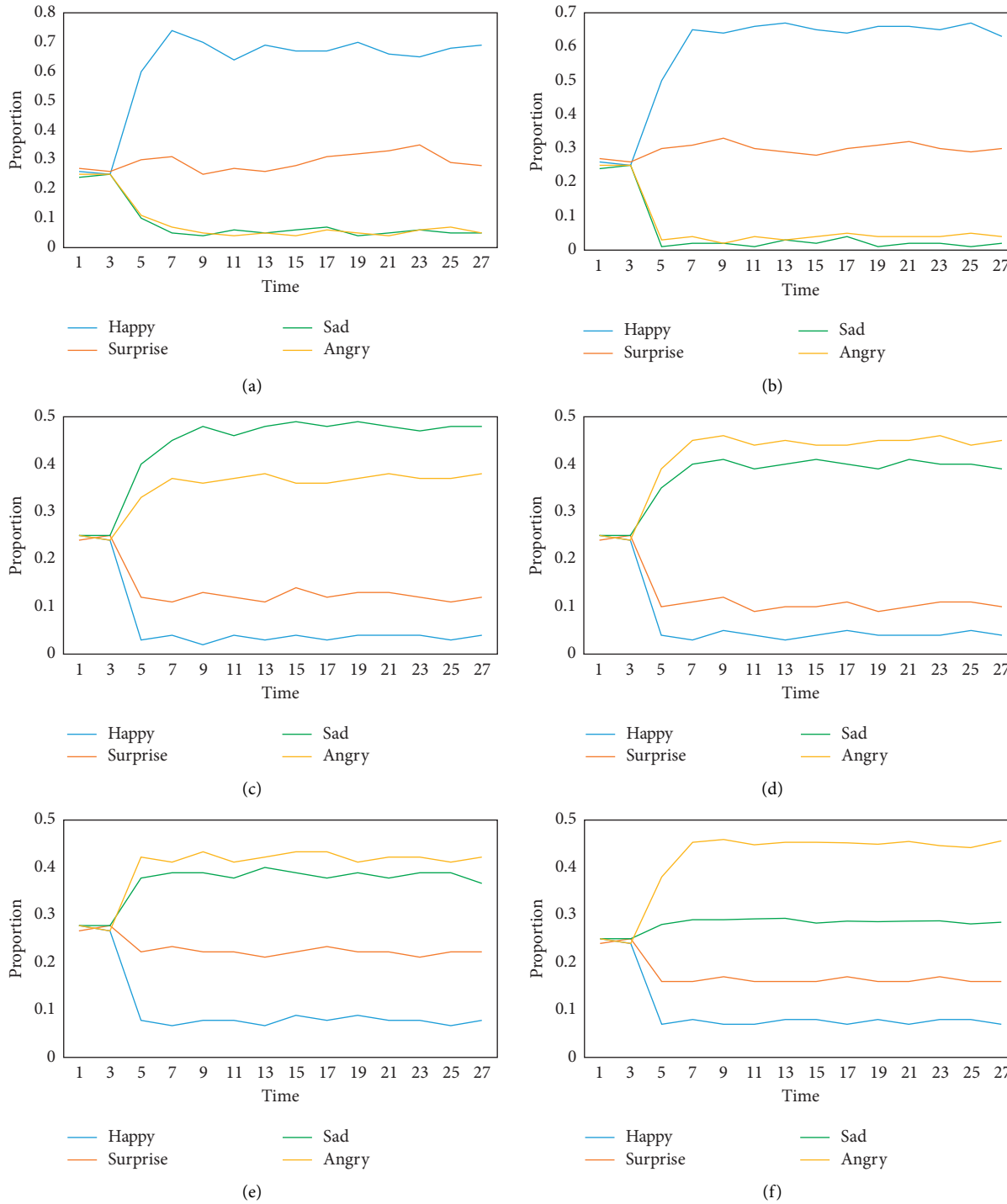

Figure 7: Continued.

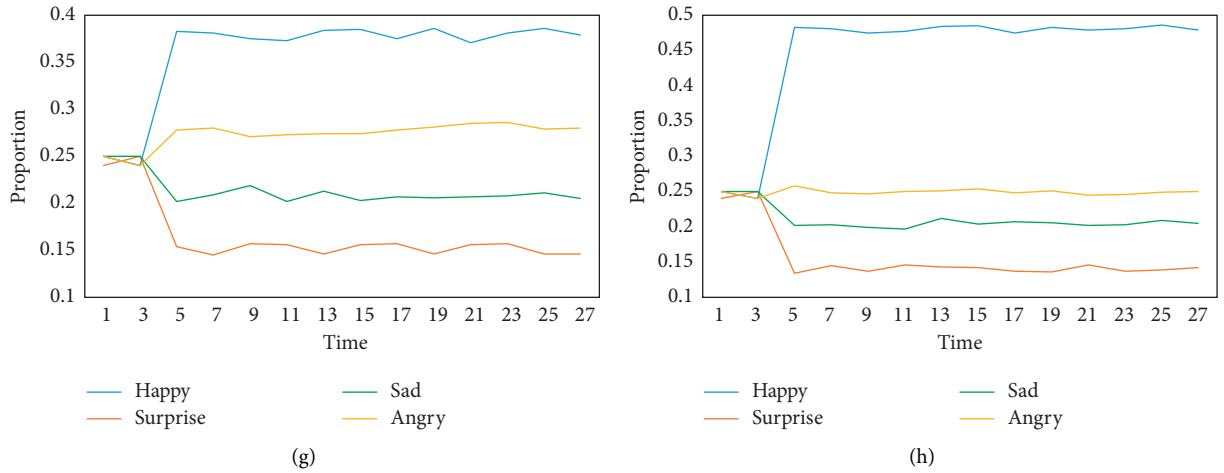

Figure 7: The changing of sentiment in four datasets. (a) Military parade in National Day topic #1. (b) Military parade in National Day topic #2. (c) The assault on a doctor topic #1. (d) The assault on a doctor topic #2. (e) Hong Kong's event topic #1. (f) Hong Kong's event topic #2. (g) Garbage sorting in Shanghai topic #1. (h) Garbage sorting in Shanghai topic #2.

value of topic #2 curve is lower than that of topic #1, which reflects the discussion of the same event will fade over time. Even if there is a new topic, discussion of a new topic is far lower than the beginning of the event. Meanwhile, similar results can be verified in other three datasets.

The proportion of the sentiment polarity in the four datasets is shown in Figure 7. Since the sentiment polarity proportion is measured, the four sentiment polarities are balanced distributed before the occurrence of the events. After the event occurred, the polarity of positive and negative sentiment began to change toward the two extremes. Among the four datasets, the positive sentiment was higher than the negative sentiment in the first dataset "National Day military parade event" and the fourth dataset "Shanghai garbage classification event," which also conforms to the social sentiment of events. In addition, it can be found that the difference among the four sentiment labels is large in the initial stage, and the distribution of sentiment labels becomes stable in the later stage. It has proved that the second report of social events does not cause the heat for the first time. But, the sentiment tendency judgment in social networks will not decline sharply with the reduction of discussion, which can be proved in topic #2 of the second dataset "The assault on a doctor." Given that the topic features come from the background of a corpus and contain a lot of noise words, the relative position of four curves is closer in terms of sentiment polarity evolution. However, there is still a gap between government feelings, which is different from the average distribution of sentiment polarity at the beginning of the event.

6. Discussion

From the perspective of theoretical significance, this paper extends the LDA model to some extent. First, in view of the sparse matrix caused by the short text, word pairs are introduced to replace a single word for text generation according to BTM. Based on the hypotheses of JST and ASUM, the sentiment layer is introduced to form the Bayesian network structure, and the word pair is limited to the same sentiment polarity distribution. Second, in order to realize dynamic analysis and text homogeneity, the timestamp and the corresponding superparameter are introduced to alleviate the problem of the word order in the text generation. Third, this research combines behavioral experiments, big data mining, mathematical modeling, and imitating to promote the research expansion of new situations and new methods.

From the perspective of practical significance, this paper is of great value in tracking and monitoring topics of public opinion in social networks. The online public opinions of hot events can be monitored, which contributes to accurately judge social events and make emergency decisions for government or departments. In addition, this paper analyzed the evolution, response, and governance of public opinion, which is conducive to understand the formation mechanism and the collaborative evolution of public opinion. Meanwhile, the use of public opinion information can detect and screen information, prevent the spread of rumor, and scientifically formulate the mechanism of utilization to effectively reduce the loss risk.

7. Conclusions

In the context of the mobile social network, the number of short texts is growing explosively. In order to extract information from massive short texts quickly and monitor public opinions, the TSTS model is proposed in the study based on LDA, BTM, JST, ASUM, and TOT. From the experimental results, the TSTS model achieves good performance. In terms of topic feature extraction, the degree of perplexity of TSTS is always lower than LDA. Moreover, although the degree of perplexity is slightly higher than BTM with the increase of iteration times, it can maintain the balance with BTM. In the sentiment analysis, the effect of TSTS was significantly better than JST and ASUM. Finally,

the TSTS model incorporating the time factor can determine the change trend of the topic and sentiment.

There are still some shortcomings in this paper. Firstly, for the extraction of topic feature words, the global topic level can be added to the topic layer of the TSTS model to filter the common topic words. Secondly, in sentiment polarity judgment, the sentiment labels are manually marked based on prior knowledge. However, the sentiments are extremely rich and changeable. In the future research, the Bayesian network and entity theory can be used to judge sentiment bias.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Project of National Science and Technology Department (2018YFF0213102), Public Projects of Zhejiang Province (LGF18G010003, LGF19G010002, and LGF20G010002), Science and Technology Project of Zhejiang Province (2020C01158), and First Class Discipline of Zhejiang - A (Zhejiang Gongshang University - Statistics).

References

[1] CNNIC, "Statistical report on internet development in China," CNNIC, Beijing, China, 2019, http://www.cac.gov.cn/2019-08/30/c_1124938750.htm.

[2] M. Ingawale, A. Dutta, R. Roy, and P. Seetharaman, "Network analysis of user generated content quality in wikipedia," *Online Information Review*, vol. 37, no. 4, pp. 602–619, 2013.

[3] Q. Gao, Y. Tian, and M. Tu, "Exploring factors influencing Chinese user's perceived credibility of health and safety information on Weibo," *Computers in Human Behavior*, vol. 45, pp. 21–31, 2015.

[4] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," *Handbook of Latent Semantic Analysis*, Psychology Press, vol. 427, no. 7, pp. 424–440, New York, NY, USA, 2007.

[5] J.-F. Yeh, Y.-S. Tan, and C.-H. Lee, "Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation," *Neurocomputing*, vol. 216, pp. 310–318, 2016.

[6] X. Zhou and L. Chen, "Event detection over twitter social media streams," *The VLDB Journal*, vol. 23, no. 3, pp. 381–400, 2014.

[7] Y. J. Du, Y. T. Yi, and X. Y. Li, "Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet allocation," *Engineering Applications of Artificial Intelligence*, vol. 87, pp. 1–13, 2020.

[8] H. J. Kim, Y. K. Jeong, Y. Kim et al., "Topic-based content and sentiment analysis of Ebola virus on twitter and in the news," *Journal of Information Science*, vol. 42, no. 6, pp. 763–781, 2016.

[9] B. Subeno, R. Kusumaningrum, and F. Farikhin, "Optimisation towards latent Dirichlet allocation: its topic number and collapsed gibbs sampling inference process," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3204–3213, 2018.

[10] H. Park, T. Park, and Y.-S. Lee, "Partially collapsed gibbs sampling for latent Dirichlet allocation," *Expert Systems with Applications*, vol. 131, pp. 208–218, 2019.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.

[13] J. Rashid, S. M. A. Shah, and A. Irtaza, "Fuzzy topic modeling approach for text mining over short text," *Information Processing & Management*, vol. 56, no. 6, pp. 1–19, 2019.

[14] H.-Y. Lu, N. Kang, Y. Li, Q.-Y. Zhan, J.-Y. Xie, and C.-J. Wang, "Utilizing recurrent neural network for topic discovery in short text scenarios," *Intelligent Data Analysis*, vol. 23, no. 2, pp. 259–277, 2019.

[15] L. Zhu, H. Xu, Y. Xu et al., "A joint model of extended LDA and IBTM over streaming Chinese short texts," *Intelligent Data Analysis*, vol. 23, no. 3, pp. 681–699, 2019.

[16] M. Tang, J. Jin, Y. Liu et al., "Integrating topic, sentiment and syntax for modeling online product reviews: a topic model approach," *Journal of Computing and Information Science in Engineering*, vol. 19, no. 1, pp. 1–12, 2019.

[17] R. K. Amplayo, S. Lee, and M. Song, "Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis," *Information Sciences*, vol. 454-455, pp. 200–215, 2018.

[18] L. Yao, Y. Zhang, B. Wei et al., "Concept over time: the combination of probabilistic topic model with Wikipedia knowledge," *Expert Systems with Applications*, vol. 60, pp. 27–38, 2016.

[19] S. Park, W. Lee, and I.-C. Moon, "Associative topic models with numerical time series," *Information Processing & Management*, vol. 51, no. 5, pp. 737–755, 2015.

[20] P. Lorenz-Spreen, F. Wolf, J. Braun et al., "Tracking online topics over time: understanding dynamic hashtag communities," *Computational Social Networks*, vol. 5, no. 1, pp. 1–18, 2018.

[21] Y. He, C. Lin, W. Gao et al., "Dynamic joint sentiment-topic model," *ACM Transactions on Intelligent Systems & Technology*, vol. 5, no. 1, pp. 1–21, 2013.

[22] L. Kuo and T. Y. Yang, "An improved collapsed Gibbs sampler for Dirichlet process mixing models," *Computational Statistics & Data Analysis*, vol. 50, no. 3, pp. 659–674, 2006.

[23] Y. Papanikolaou, J. R. Foulds, T. N. Rubin et al., "Dense distributions from sparse samples: improved Gibbs sampling parameter estimators for LDA," *Statistics*, vol. 18, no. 62, pp. 1–58, 2015.

[24] X. Zhou, J. Ouyang, and X. Li, "Two time-efficient Gibbs sampling inference algorithms for biterm topic model," *Applied Intelligence*, vol. 48, no. 3, pp. 730–754, 2018.

[25] P. Bhuyan, "Estimation of random-effects model for longitudinal data with non ignorable missingness using Gibbs sampling," *Computational Statistics*, vol. 34, no. 4, pp. 1963–1710, 2019.

[26] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 1134–1145, 2012.

[27] A. Daud and F. Muhammad, "Group topic modeling for academic knowledge discovery," *Applied Intelligence*, vol. 36, no. 4, pp. 870–886, 2012.

[28] Z. Huang, J. Tang, G. Shan, J. Ni, Y. Chen, and C. Wang, "An efficient passenger-hunting recommendation framework with multi-task deep learning," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7713–7721, 2019.

[29] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013)*, Springer, Atlanta, GA, USA, pp. 1–5, June 2013.

[30] T. Chen, Q. Li, J. Yang, G. Cong, and G. Li, "Modeling of the public opinion polarization process with the considerations of individual heterogeneity and dynamic conformity," *Mathematics*, vol. 7, no. 10, p. 917, 2019.

[31] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: twitter and Reddit," *Information Processing and Management*, vol. 57, no. 2, pp. 1–21, 2019.