International Journal of
Geo-Information

Article

Topic-Clustering Model with Temporal Distribution for Public Opinion Topic Analysis of Geospatial Social Media Data

Chunchun Hu [1], Qin Liang [2,*], Nianxue Luo [1] and Shuixiang Lu [2]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430072, China; chchhu@sgg.whu.edu.cn (C.H.)
[2] Zhejiang Academy of Surveying and Mapping, Hangzhou 311100, China
* Correspondence: zjasm_lq@163.com

Abstract: Analysis of the spatiotemporal distribution of online public opinion topics can help understand the hotspots of public concern. The topic model is employed widely in public opinion topic clustering for social media data. In order to handle topic-clustering of low-quality geospatial social media data, such as microblog data, with short text and timeliness characteristics, this study proposed a Dirichlet multinomial mixture over time (DMMOT) model to cluster microblog topic for public opinion analysis. The DMMOT model assumes that a single document belongs to a single topic, in line with the characteristics of a short text, and it introduces the probability distribution of "topic-time" in the process of topic generation. The model parameter inference process was presented in detail by exploring the Gibbs sampling method. Results generated using the DMMOT model in case study show that the "topic-word" distribution is semantically aggregated within various topics, and "topic-time" distribution clustered within a time window under each topic. Furthermore, the characteristics of the trend of each topic over time are basically consistent with the corresponding trend of topic in reality in terms of content. These indicate that the DMMOT model improves topic clustering for short text to some extent. Furthermore, the DMMOT model performed well in both temporal and spatial analysis of public opinion topics based on microblog data.

Keywords: topic clustering; microblog data; DMMOT model; public opinion analysis; Gibbs sampling

check for updates

Citation: Hu, C.; Liang, Q.; Luo, N.; Lu, S. Topic-Clustering Model with Temporal Distribution for Public Opinion Topic Analysis of Geospatial Social Media Data. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 274. https://doi.org/10.3390/ijgi12070274

Academic Editors: Wolfgang Kainz and Wei Huang

Received: 30 March 2023
Revised: 16 June 2023
Accepted: 29 June 2023
Published: 8 July 2023

1. Introduction

Social media amass rich user-generated content such as text, pictures, locations, and videos. Such abundant data provide a new perspective to study the spatiotemporal behavior characteristics of humans and reveal the spatial distribution patterns and spatiotemporal evolution process of social phenomena [1]. When a significant event occurs, the public tends to post by reflecting its attention to and cognition of the event on social media platforms such as Weibo, thus triggering the topic's dissemination and discussion. Social media gather a large amount of semantic data, which have become an important source to understand social situations and public opinion [2]. Such data have been analyzed using text mining and various geospatial methods to determine the public's behavior over time and their related opinions in a specific environment.

Further, because of the increasing popularity of social media research, many scholars have used data from social media to analyze public opinion in the field of emergencies, such as natural disasters and public health [3–8]. With the COVID-19 (Corona Virus Disease 2019) outbreak, social media quickly became an important data source for information generation and dissemination. Further, it is of great practical significance to objectively understand public opinion and regional differences during the pandemic to improve the policy regulation and scientific governance of major public health events.

Sina Weibo is an important social media platform for online users to record their lives, express and exchange opinions, and obtain and share information. Weibo has a large user base. As of September 2020, the monthly active users have reached 511 million, and the daily active users have reached 224 million [9]. Therefore, Weibo platform has become an important source of information acquisition, and microblog data obtained from Weibo is widely used in various studies such as behavior analysis, interest recommendation, urban planning, public safety, and social perception. This study aims to mine and analyze public opinion on an urban scale from microblog data by exploring a topic-clustering model. The latent Dirichlet allocation (LDA) model is one of the most popular topic models, with the advantage of automatically identifying semantic topics from massive texts [10]. The traditional LDA topic model mainly discovers latent topics based on word co-occurrence. Because of the lack of co-occurrence information between the words in short texts, the LDA model is not effective enough when applied to short texts. Instead, the Dirichlet multinomial mixture (DMM) [11] model is more suitable for short text topic modelling as it assumes that each document is generated from one topic. Moreover, the topics over time (TOT) [12] model uses beta distribution to fit the distribution of topics over time to discover the topics' evolution processes. However, it is difficult to distinguish between different topics if the words show little difference in their temporal distribution.

Accordingly, this study proposes a new topic-clustering model—the Dirichlet multinomial mixture over time (DMMOT) model—which combines the advantages of the DMM and TOT models to handle low-quality microblog data with short text and timeliness characteristics. The hypothesis conditions of the DMMOT model are more in line with the characteristics of microblog data, and this model can directly identify the temporal distribution probability of topics and the unique topic to which a document belongs. This research can help obtain the spatiotemporal evolution characteristics of public opinion topics about hot social issues.

2. Related Works

The topic model was originally a probabilistic model proposed by Hofmann [13] based on latent semantic indexing. Blei et al. [14] added the probability distribution and parameters' prior information to the probabilistic latent semantic indexing model to obtain a more complete probability generation model called the LDA topic model. In topic models, a topic is added as a hidden variable to the generative model of documents and words, and LDA model is the basis for all the extended topic models.

Most extended probabilistic topic models redefined the assumptions of the document generation process to solve the limitation of existing topic models; then, latent topics were derived based on the results of parametric statistical inference. Blei et al. [15] changed the Dirichlet prior distribution of the LDA model to a logistic-normal distribution and obtained a correlated topic model by increasing the correlations between latent topics, and found that an estimated Correlated Topic Model (CTM) can be used to explore otherwise unstructured observed documents. Li et al. [16] obtained the labeled-LDA model by building the documents' category levels, which weakened the phenomenon wherein topics that did not belong were forcibly assigned to certain documents in the LDA model.

With the rise of social media, recent research mainly focused on solving the problem of data sparsity in documents and topic models for short texts. Further, researchers of [17] have improved the problem of text sparsity by changing the model structure with the hypothesis condition of model generation. The DMM [11] model is a typical topic model for short texts that assumes independence between the words in one document, which dealt with the sparse and high-dimensional problem of short texts, and can obtain the representative words of each cluster. Ma et al. [18] aggregated short texts into long texts according to related characteristics, which introduced external knowledge to assist document clustering using a topic model. To handle short documents well, the biterm topic model (BTM) [17] introduced the biterm, which is an unordered co-occurrence word pair in a short document, based on the LDA model. The BTM model assumes that the topic of each word also depends on the previous word, thus contributing to the expansion of document length. However, the BTM model considers word order and is still influenced by the sparsity of word order. A previous study [19] showed that short-distance co-occurrence information is more valuable than long-distance co-occurrence information, and increasing the window length contributes little to the number of co-occurrence words. Relevant study [20] has introduced a word-embedding model based on these models and added semantic information. The local LDA model [21] focused on short distance co-occurrence but did not assume inter-word dependencies. It added a fixed size overlapping window to the LDA model to extract more coherent and meaningful topics. The above research attempts to solve the problem of data sparsity of short text clustering from various aspects, but they do not pay attention to the sequential variation characteristics of short texts.

The probabilistic models with time series were developed to analyze the time evolution of topics in large document collections. Current temporal topic models were proposed to simulate and infer the evolutionary trends of topics. For instance, the dynamic topic model [22] considers discrete time in the topic model, but the model results are affected by the size of the time window. However, for latent topics in the document, the TOT [12] model fits the distribution of topics over time through beta distribution, which solved problem caused by time window. These studies indicate that the results generated by topic model can be dynamically varied. Specially, for social media data, the topic clustering would be influenced by time information in addition to word co-occurrences in text.

In addition, the parameter estimation method for topic models mainly includes Variational Bayesian (VB) inference and Gibbs sampling [18]. The inference process of VB is complicated and related to setting of the initial parameter values. It may yield local optimal results and has high time efficiency. However, the mathematical derivation process of Gibbs sampling is simpler than the VB inference and is less affected by the parameters' initial values. Collapsed Gibbs sampling uses the Bayesian formula to solve the integral of the hidden variables in the model to eliminate unknown parameters. Thus, it can overcome the problems caused by the inference using approximate posterior distributions and facilitate more accurate parametric statistical inference of the model. Thus, this article adopted Gibbs sampling method to conduct parameter inference.

Prior research on topic models mainly focused on the analysis of online public opinion [23–26]. For instance, Han [23] combined the LDA topic and random forest models to obtain different topic contents of microblog texts related to the pandemic situation as well as obtain the spatiotemporal distribution characteristics of each public opinion topic on a national scale. Han found that the change trend of the topic-time of public opinion and the pandemic development trend is synchronized, while the topic's spatial distribution is related to pandemic severity, population intensity, and so on. Wang [24] built a topic extraction and classification model to analyze the distribution characteristics of public opinion in focal areas, such as several metropolitan areas and border ports in China. Boon-Itt and Skunkan [25] conducted sentiment analysis and LDA topic modeling based on Twitter data, showing that the public was mainly pessimistic during the outbreak. They revealed three pandemic-related topics: pandemic emergency, pandemic control, and pandemic reporting. Amara et al. [26] used a multilingual corpus from Facebook to extract topics using LDA topic modeling to track trends of the COVID-19 pandemic. In the above applications, LDA topic model was widely adopted to analyze the public opinion, and the additional information in datasets was not incorporated to improve the model while LDA model may not be the optimal solution.

3. Topic-Clustering Model with Temporal Distribution

In the current context, although Weibo has removed the 140-character limit, most microblogs have fewer words. Thus, there is less word co-occurrence information in microblog text. When applying the LDA model to the analysis of such short texts in social media, the generated results might be of poor quality because of less word co-occurrence information in such short texts. To meet the needs of short texts, the DMM model assumes that one document belongs to only one topic, and all the words under the document stem from that one topic. By considering the topic evolution processes, the TOT model incorporates temporal features based on LDA, fitting the temporal distribution features of document topics using a beta distribution.

To make better use of low-quality microblog data, this study integrates the user's check-in time to build the DMMOT model by combining the temporal evolution characteristics with the DMM model for topic clustering of microblog text; we also explore the Gibbs sampling method for parametric inference of our model.

3.1. DMMOT Topic Model

The DMMOT is a probabilistic generative document model. It assumes that one document belongs to only one topic, and all words in the document are independent of each other. Document topics need to be sampled only once. All words in a document are sampled from one topic once the document's topics are assigned, and both "topic-word" and "document-topic" distributions behave according to a polynomial distribution. For microblog texts, different topics maintain high prevalence for a certain period, and then, this prevalence slowly disappears over time. With reference to the literature [12], the temporal distribution features of topics can be fitted using beta distribution. The probability density function of the beta distribution is given by Equation (1).

$$f(x;\psi_1,\psi_2) = \frac{\Gamma(\psi_1+\psi_2)}{\Gamma(\psi_1)\Gamma(\psi_2)}x^{\psi_1-1}(1-x)^{\psi_2-1} = \frac{x^{\psi_1-1}(1-x)^{\psi_2-1}}{B(\psi_1,\psi_2)}, \tag{1}$$

where Γ is the gamma function, and $B(\psi_1,\psi_2)$ is the beta function.

The DMMOT model is suitable for short texts, such as microblog text. Figure 1 shows this model's generation process where the box represents the number of repeated samplings, the circles with shadows are observable variables, and the other circles are unobservable variables. In the figure, D is the number of documents, N_d is the number of words included in document d, and K is the number of topics. The DMMOT model selects the document's topics from the polynomial distribution of document-topic and then selects words under the chosen topics from the polynomial distribution of topic-word to yield the document. Here, α is the hyperparameter of the Dirichlet prior distribution of the document-topic distribution, and β is the hyperparameter of the Dirichlet prior distribution of the topic-word distribution.

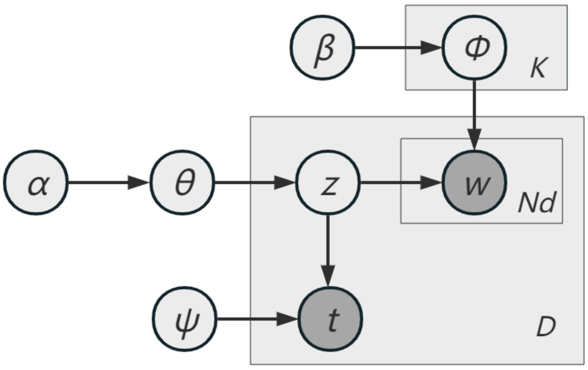


Figure 1. Dirichlet multinomial mixture over time model.

Document generation using the DMMOT model is divided into the following three steps:

I. Select topic z of document d according to polynomial distribution θ [$\theta \sim Dir(\alpha)$] of the document-topic;

II. Select the word distribution under the zth topic according to polynomial distribution Φ [$\Phi \sim Dir(\beta)$] of the topic-word. Then, N_d words of the document are obtained by sampling N_d times under the probability distribution;

III. According to the time attribute of documents, the value of probability that one document belongs to a different topic is obtained from the probability distribution of "topic-time," which is part of the Gibbs sampling probability.

3.1.1. Inference Process of the DMMOT Model by Gibbs Sampling

Gibbs sampling is a special Markov chain algorithm. The Markov chain assumes that the transition probability of the current state is only related to the previous state. When setting the arbitrary probability distribution of the initial state, distribution π converges to a stationary distribution after the state transition matrix P of the Markov chain model transforms. Therefore, when the probability distribution of the sample is unknown, an arbitrary initial probability distribution can be set. Given a Markov chain, distribution π, and probability transition matrix P, if the following equation holds, satisfying the detailed balance condition, then this Markov chain has a stationary distribution π:

$$\pi(i)P(i,j) = \pi(j)P(j,i), \tag{2}$$

Gibbs sampling iteratively converges to the probability stationary distribution by sampling only one dimension when fixing the other dimensions every time, while taking the conditional probability of stationary distribution π as the sampling probability. This can yield sampling results based on the joint probability density calculated by sampling based on the conditional probability density function. Therefore, when applying Gibbs sampling to topic modeling, it is necessary to first obtain the joint probability density generated from the corpus and derive the conditional probability formulation for topic sampling.

In the DMMOT model, the Dirichlet distribution serves as a conjugate prior to the polynomial distribution. Table 1 describes the parameters in the DMMOT topic model. If parameter θ follows a Dirichlet distribution, it is denoted as $\theta|\ \alpha \sim Dir(\alpha)$. The document, words, and time in the DMMOT model obey the probability distributions in Equation (4):

$$\theta|\,\alpha \sim Dir(\alpha);\ \Phi|\beta \sim Dir(\beta), \tag{3}$$

$$z|\,\theta \sim Mult(\theta);\ w|\Phi \sim Mult(\Phi);\ t|\psi \sim Beta(\psi) \tag{4}$$

Table 1. Representation of the parameters in the DMMOT model.

Parameter	Meaning of Parameters
m_k	Number of documents belonging to topic k
D	Number of documents in the corpus
K	Topic number
V	Number of words in the corpus
α, β	Hyperparameter of Dirichlet prior distribution ofthe "document-topic" and "topic-word"
N_d^w	Number of times word w appears in document d
N_d	Total number of words in document d
n_k^w	Number of times of word w belongs to topic k
n_k	Total number of words belonging to topic k
t_d	Time attribute value for document d
$\psi_{z_d}1, \psi_{z_d}2$	Two parameters of the beta distribution for document z_d

To use the Gibbs sampling method, joint probability of the generated document is first obtained, as shown in Equation (5):

$$\begin{gathered} p\left(\vec{d}, \vec{z}, \vec{t}\,\middle|\,\alpha, \beta, \psi\right) \\ = p(d|z,\beta)p(z|\alpha)p(t|z,\psi) \\ = \int\ p(d|z,\Phi)p(\Phi|\beta)d\Phi \int\ p(z|\theta)p(\theta|\alpha)d\theta p(t|z,\psi) \end{gathered} \tag{5}$$

The following equations can be derived from the probability density function of the probability distribution of document-topic, topic-word, and topic-time:

$$\begin{gathered}\int p(d|z,\Phi)p(\Phi|\beta)d\Phi \\ =\int \prod_{k=1}^{K}\prod_{t=1}^{V}\Phi_{k,t}^{n_k^t}\prod_{k=1}^{K}\frac{1}{\Delta(\beta)}\prod_{t=1}^{V}\Phi_{k,t}^{\beta_t-1}d\Phi \\ =\int \prod_{k=1}^{K}\frac{1}{\Delta(\beta)}\prod_{t=1}^{V}\Phi_{k,t}^{n_k^t+\beta_t-1}d\Phi \\ =\prod_{k=1}^{K}\frac{1}{\Delta(\beta)}\int \prod_{t=1}^{V}\Phi_{k,t}^{n_k^t+\beta_t-1}d\Phi \\ =\prod_{k=1}^{K}\frac{\Delta(n_k+\beta)}{\Delta(\beta)},\end{gathered} \tag{6}$$

where $\Delta(\beta)=\int \prod_{i=1}^{V}p_i^{\beta_i-1}dp=\frac{\prod_{i=1}^{V}\Gamma(\beta_i)}{\Gamma\left(\sum_{i=1}^{V}\beta_i\right)}$, $n_k=\sum_{t=1}^{V}n_k^t$,

$$\begin{aligned}\int p(z|\theta)p(\theta|\alpha)d\theta &=\int \prod_{k=1}^{K}\theta_k^{m_k}\frac{1}{\Delta(\alpha)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}d\theta \\ &=\frac{1}{\Delta(\alpha)}\int \prod_{k=1}^{K}\theta_k^{m_k+\alpha_k-1}d\theta \\ &=\frac{\Delta(m+\alpha)}{\Delta(\alpha)},\end{aligned} \tag{7}$$

where $\Delta(\alpha)=\int \prod_{i=1}^{K}p_i^{\alpha_i-1}dp=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$, $m=\sum_{k=1}^{K}m_k$,

$$p(t|z,\psi)=\prod_{d=1}^{D}p\left(t_d\middle|\psi_{z_d}\right) \tag{8}$$

Combining Equations (6)–(8), the joint probability in Equation (5) can be defined as follows:

$$\begin{gathered}p\left(\vec{d},\vec{z},\vec{t}\middle|\alpha,\beta,\psi\right) \\ =\prod_{k=1}^{K}\frac{\Delta(n_k+\beta)}{\Delta(\beta)}\times\frac{\Delta(m+\alpha)}{\Delta(\alpha)}\times\prod_{d=1}^{D}p\left(t_d\middle|\psi_{z_d}\right)\end{gathered} \tag{9}$$

According to the Bayesian formula of joint probability, the conditional probability that document d belongs to topic k is defined as follows:

$$\begin{gathered}p\left(z_d=k\middle|\vec{d},z_{-d},\vec{t},\alpha,\beta,\psi\right) \\ =\frac{p\left(d,k,t\middle|\vec{d}_{-d},\vec{z}_{-d},\vec{t}_{-d},\alpha,\beta,\psi\right)}{p\left(d,t\middle|\vec{d}_{-d},\vec{z}_{-d},\vec{t}_{-d},,\alpha,\beta,\psi\right)} \\ \propto\frac{p\left(\vec{d},\vec{z},\vec{t}\middle|\alpha,\beta,\psi\right)}{p\left(\vec{d}_{-d},\vec{z}_{-d},\vec{t}_{-d}\middle|\alpha,\beta,\psi\right)} \\ \propto\frac{\Delta(n_k+\beta)}{\Delta\left(n_{k,-d}+\beta\right)}\frac{\Delta(m+\alpha)}{\Delta(m_{-d}+\alpha)}p\left(t_d\middle|\psi_{z_d}\right)\end{gathered} \tag{10}$$

Based on the continuous multiplication property of the gamma function, we further derive Equations (11) and (12):

$$\begin{aligned}\frac{\Delta(n_k+\beta)}{\Delta\left(n_{k,-d}+\beta\right)} &= \frac{\Gamma(n_k+\beta)}{\Gamma\left(n_{k,-d}+\beta\right)}\frac{\Gamma\left(\sum_{t=1}^{V}\left(n_{t,-d}+\beta\right)\right)}{\Gamma\left(\sum_{t=1}^{V}(n_t+\beta)\right)} \\ &= \frac{\prod_{w\in d}\Gamma\left(n_k^w+\beta\right)}{\prod_{w\in d}\Gamma\left(n_{k,-d}^w+\beta\right)}\frac{\Gamma\left(n_{k,-d}+V\beta\right)}{\Gamma(n_k+V\beta)} \\ &= \frac{\frac{\prod_{w\in d}\Gamma\left(n_k^w+\beta\right)}{\prod_{w\in d}\Gamma\left(n_{k,-d}^w+\beta\right)}}{\prod_{i=1}^{N_d}\left(n_{k,-d}+V\beta+i-1\right)},\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\Delta(m+\alpha)}{\Delta(m_{-d}+\alpha)} &= \frac{\Gamma(m+\alpha)}{\Gamma(m_{-d}+\alpha)}\frac{\Gamma\left(\sum_{k=1}^{K}\left(m_{k,-d}+\alpha\right)\right)}{\Gamma\left(\sum_{k=1}^{K}(m_k+\alpha)\right)} \\ &= \frac{\Gamma(m_z+\alpha)}{\Gamma\left(m_{z,-d}+\alpha\right)}\frac{\Gamma(D-1+K\alpha)}{\Gamma(D+K\alpha)} \\ &= \frac{m_{z,-d}+\alpha}{D-1+K\alpha}.\end{aligned} \tag{12}$$

Based on the derivation of DMM [10], beta distribution in Equation (13), and the aforementioned formulations, we derive the sampling probability in Equation (14):

$$p\left(t_d\middle|\psi_{z_d}\right) = \frac{{t_d}^{\psi_{z_d1}-1}(1-t_d)^{\psi_{z_d2}-1}}{B\left(\psi_{z_d1},\psi_{z_d2}\right)}, \tag{13}$$

$$\begin{gathered}p\left(z_d = k\middle|\vec{d},z_{-d},\vec{t},\alpha,\beta,\psi\right) \\ \propto \frac{\frac{\prod_{w\in d}\Gamma\left(n_k^w+\beta\right)}{\prod_{w\in d}\Gamma\left(n_{k,-d}^w+\beta\right)}}{\prod_{i=1}^{N_d}\left(n_{k,-d}+V\beta+i-1\right)} \times \frac{m_{k,-d}+\alpha}{D-1+K\alpha} \times \frac{{t_d}^{\psi_{z_d1}-1}(1-t_d)^{\psi_{z_d2}-1}}{B\left(\psi_{z_d1},\psi_{z_d2}\right)}.\end{gathered} \tag{14}$$

In Equation (14), parameters of the beta distribution, ψ_{z_d1}, ψ_{z_d2} can be computed using the moment estimation method. The mean value $\bar{t}_z$ and variance s_z^2 can be computed as follows:

$$\bar{t}_z = \frac{\psi_{z_d1}}{\psi_{z_d1}+\psi_{z_d2}}, \tag{15}$$

$$s_z^2 = \frac{\psi_{z_d1}\psi_{z_d2}}{\left(\psi_{z_d1}+\psi_{z_d2}\right)^2\left(\psi_{z_d1}+\psi_{z_d2}+1\right)}. \tag{16}$$

According to the mean and variance of the beta distribution of topic z in Equations (15) and (16), parameters of the beta distribution can be obtained as follows:

$$\psi_{z_d1} = \bar{t}_z\left(\frac{\bar{t}_z\left(1-\bar{t}_z\right)}{s_z^2}-1\right), \tag{17}$$

$$\psi_{z_d2} = \left(1-\bar{t}_z\right)\left(\frac{\bar{t}_z\left(1-\bar{t}_z\right)}{s_z^2}-1\right). \tag{18}$$

3.1.2. Gibbs Sampling Algorithm for the DMMOT Model

According to the sampling probability in Equation (14) of the DMMOT model, we obtain the posterior distribution of the document-topic, topic-word, and topic-time by constantly assigning topics to all documents in the corpus during each iteration. Specifically, the definition domain of the beta function is between 0 and 1, and temporal properties of the documents must be normalized before sampling. Details of the Gibbs sampling algorithm are shown in Algorithm 1.

Algorithm 1: The Gibbs Sampling algorithm for DMMOT model

```
Input: the number of topics $K$, the number of document $D$ and number of iterations $iter$
Output: the topic classification labels $K_d$ for all documents
set up hyper-parameters of Dirichlet distribution $\alpha,\ \beta$, let $m_k,\ n_k,\ n_k^w$ equal to 0 respectively;
for document $d \in [1,\ D]$ do
    random sampling for document $d$'s topic classification label $k_d$;
    insert $k_d$ into $K_d$;
    $m_k \leftarrow m_k + 1;\ n_k \leftarrow n_k + N_d\,;$
    for word $w \in d$ do
        $n_k^w \leftarrow n_k^w + N_d^w\,;$
    end for
end for
calculate the Mean and Variance of timestamps under every topics;
employ the method of Moments to estimate the parameter $\psi$ of Beta distribution;
for iteration $i \in [1,\ iter]$ do
    for document $d \in [1,\ D]$ do
        record the topic classification label $k = k_d$ and timestamp $t_d$ of this document $d$;
        $m_k \leftarrow m_k - 1;\ n_k \leftarrow n_k - N_d\,;$
        for word $w \in d$ do
            $n_k^w \leftarrow n_k^w - N_d^w\,;$
        end for
        sample a new topic label $k_d$ for document $d$ based on the deduced Gibbs Sampling equation;
        insert $k_d$ into $K_d$;
        $m_k \leftarrow m_k + 1;\ n_k \leftarrow n_k + N_d\,;$
        for word $w \in d$ do
            $n_k^w \leftarrow n_k^w + N_d^w\,;$
        end for
    end for
    utilize the Mean and Variance of timestamps to estimate Beta distribution parameters $\psi$ for different topics;
end for
```

4. Case Study and Discussion

The study area includes 13 municipal districts of Wuhan city. Hu et al. [27,28] pooled all posts on Weibo of every user and built a rich Weibo corpus with a user pool of 20 million active users. Based on Weibo corpus in [28], we filtered the check-in data from Wuhan City before and after the COVID-19 outbreak in Wuhan from December 2019 to April 2020. The available attributes of the data include posting time, location marked as latitude and longitude, and text.

In this study, we transformed the coordinates of the acquired Weibo data (check in microblogs), unified the coordinate system, extracted the microblog check-in data for Wuhan according to the spatial scope, and obtained 617,032 microblogs posted by 124,281 users on Weibo. Simultaneously, we counted the number of Chinese characters in microblogs and found that most users posted short texts as shown in Figure 2. After text filtering using keywords and dividing using stop words with high frequency, 46,774 non-empty microblogs were finally extracted after replacing the synonyms.

4.1. Comparison Results for the DMM and LDA Topic Models

To explore the topic of public opinion during the pandemic, this study used the DMM and LDA models for topic mining of microblog texts and compared the application of the two models to a complex text corpus. As the text topics were unknown, the number of topics for all the models was set to 15 during the experiment. Table 2 presents the meaningful topics selected from the mining results of the two models. Partial results of the DMM and LDA models are similar. Both models mined the following topics: case reports (topic 0), work resumption (topic 1), global outbreak (topic 2), praying for the end of the pandemic (topic 3), community life (topic 4), daily pandemic prevention and control (topic 5), hospital diagnosis (topic 6), treatment (topic 7), nucleic acid test (topic 8), family care (topic 9), and salute to medical staff (topic 10).

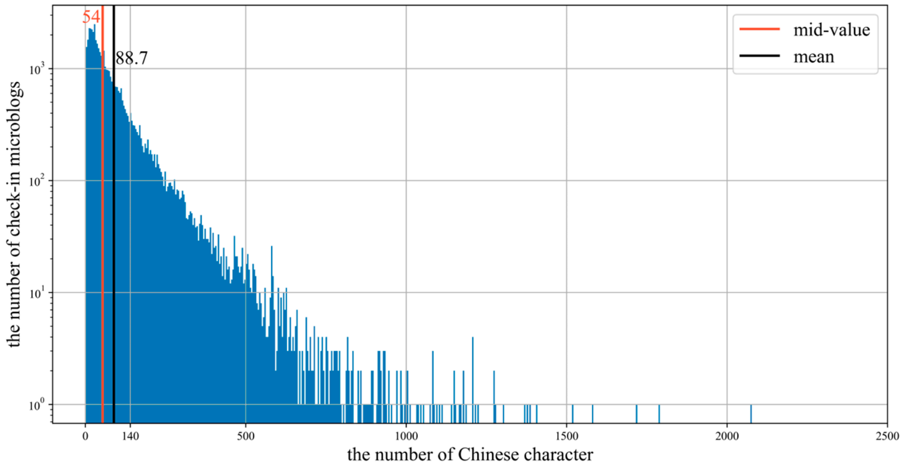


Figure 2. Statistics result for the character count of the microblogs.

Table 2. Topic-clustering results of the DMM and LDA models.

Topic ID	Words Describing the Topic of the DMM Model	Words Describing the Topics of the LDA Model
0	New, case, confirm, death, cumulative, COVID-19, number, pneumonia, suspected case, and discharge	Work resumption, new, case, confirm, firm, lift the lockdown, prevent and control, work, on duty, proof, enterprise, recover, staff, cumulative, and inform
1	Work resumption, firm, enterprise, work, influence, life, consumption, economy, express delivery, and situation	
2	Virus, China, COVID-19, pneumonia, America, infect, country, patient, coronavirus, and global	China, COVID-19, virus, country, America, global, pneumonia, world, economy, and influence
3	Lockdown, city, life, safe, early, spring, day, work resumption, Sakura, and expect	Life, safe, day, early, health, lockdown, hurry up, mood, at home, and expect
4	Neighborhood, community, estate, volunteer, resident, groupon, quarantine, confirm, supply, and go out	Neighborhood, community, volunteer, resident, estate, groupon, supply, staff, worker, and proprietor
5	Mask, go out, neighborhood, work resumption, supermarket, on duty, at home, go home, disinfect, and lockdown	Mask, go out, neighborhood, supermarket, disinfect, protection, at home, on the road, alcohol, and downstairs
6	Hospital, patient, protective clothing, mask, work, medical workers, frontline, supply, appreciate, and support	Hospital, patient, quarantine, test, nucleic acid, doctor, community, therapy, examine, CT, discharge, infect, fever, situation, and heat
7	Hospital, patient, quarantine, community, confirm, nucleic acid, pneumonia, doctor, test, and infect	
8	Work resumption, test, nucleic acid, quarantine, neighborhood, prevent and control, staff, community, lift the lockdown, and proof	
9	Mom, dad, hospital, at home, quarantine, worry, go home, go out, child, and infect	Mom, dad, on duty, protective clothing, child, go home, work, go back, at home, and husband
10	Hero, China, anti-epidemic, people, appreciate, salute, frontline, national, pneumonia, and fight	Frontline, appreciate, medical workers, anti-epidemic, people, hero, China, support, national, and fight

Topics generated by the LDA model were a mixture of multiple words that may belong to two topic categories. For example, topics 0 and 1 were clustered into one topic in the LDA model. In the DMM model, the same word appeared frequently for various topics, such as the words “hospital” and “work resumption”. This is because one document in the corpus belonged to a single topic, which was determined by the assumptions of the DMM model, and the same words may appear frequently in documents under different topics. Despite only a few instances of the same words appearing in different topics generated by the LDA model, there was low semantic coupling of words within a topic, and different categories of words may be mixed within the same topic. In contrast, each topic mined by the DMM model was semantically compact between words so that the results can express more details. For instance, hospital supplies (topic 6), patient infection and diagnosis (topic 7), and nucleic acid testing proof required for work resumption (topic 8) were clustered into one topic in the LDA model. Instead, three topics were generated for the above by the DMM model for microblog text.

Simultaneously, the DMM model is theoretically able to adaptively determine the number of topics by gradually iterating and discarding unpopular topics with small numbers of documents. However, this is related to the value α of the prior distribution.

4.2. Results of Comparing the TOT and LDA Topic Models

The TOT model adds the influence of time information on topic clustering to the LDA model. Based on the microblog corpus, the number of topics was set to 15 to compare the results of the two models. Tables 3 and 4 show the meaningful topics mined by the TOT and LDA models and their distributions over time, respectively.

Table 3. Topic-clustering results of the topics over time (TOT) topic model.

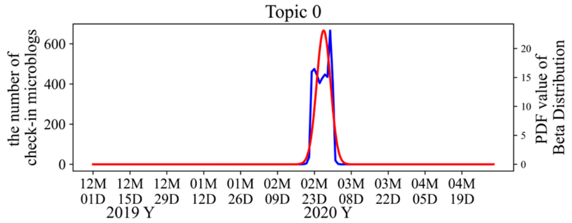
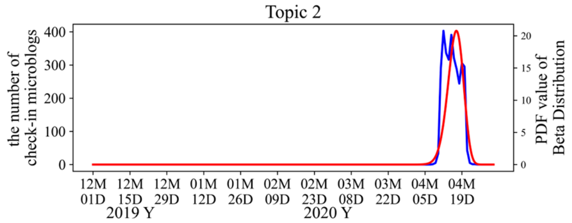
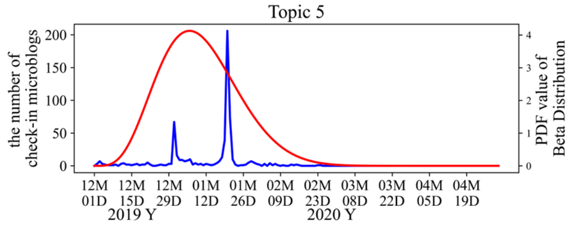
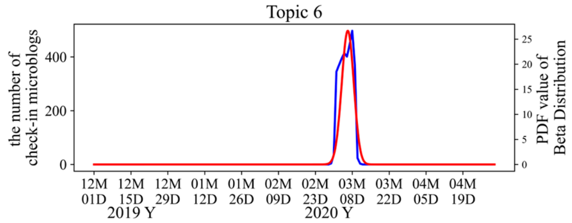
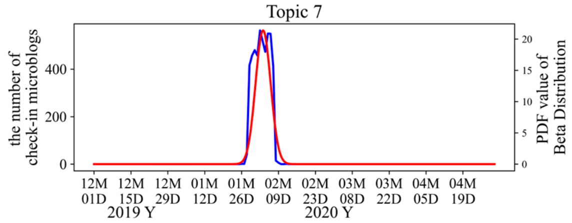
Topic ID	Words Describing the Topics of the TOT Model	Popularity of Each Topic over Time
0	Community, neighborhood, patient, quarantine, hospital, life, work, lockdown, COVID-19, and virus	
2	Work resumption, mask, test, COVID-19, nucleic acid, neighborhood, China, life, confirm, and go out	
5	Mask, pneumonia, go out, unknown, reason, vaccine, coronavirus, virus, patient, and novel	
6	Patient, hospital, community, neighborhood, life, quarantine, work, appreciate, COVID-19, and lockdown	
7	Hospital, quarantine, mask, patient, go out, confirm, infect, pneumonia, at home, and community	

Table 3. *Cont.*

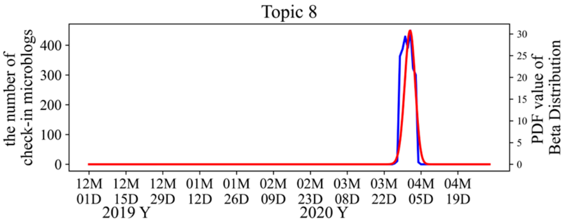
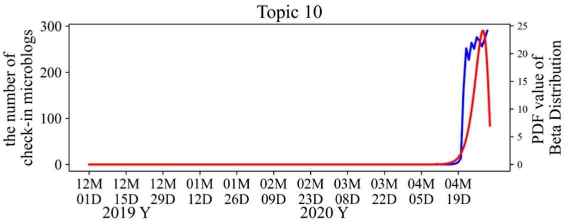
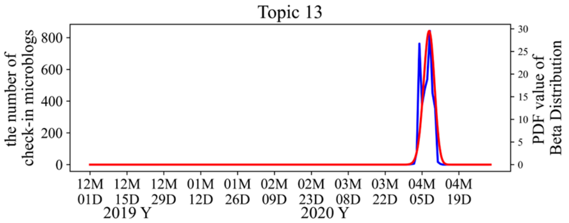
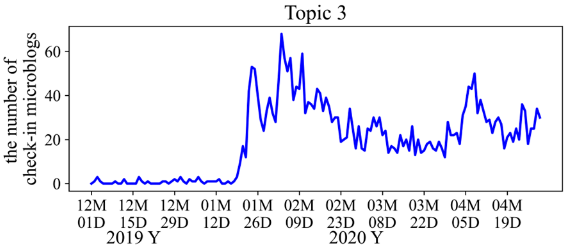
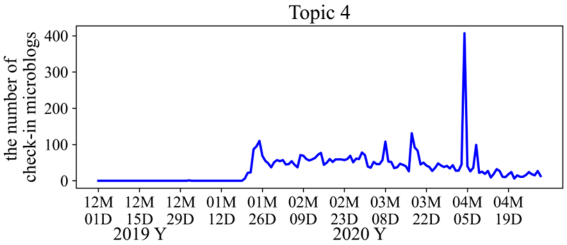
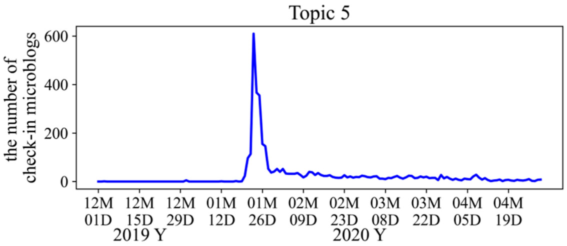
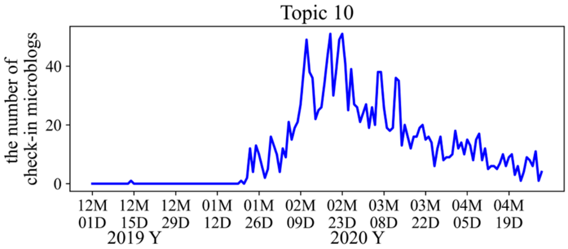
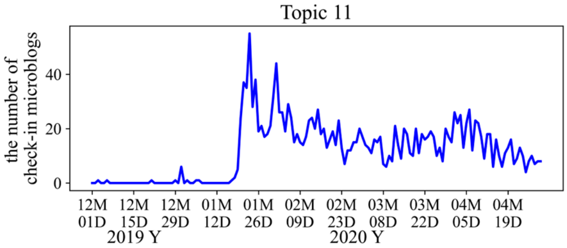
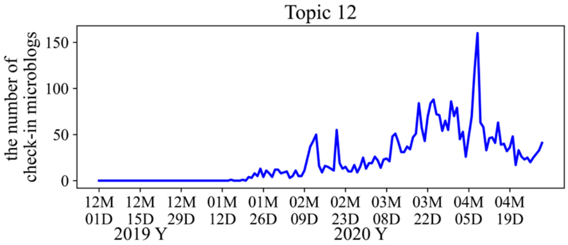
Topic ID	Words Describing the Topics of the TOT Model	Popularity of Each Topic over Time
8	Work resumption, neighborhood, virus, mask, COVID-19, quarantine, China, hospital, go out, and life	
10	Mask, work resumption, COVID-19, test, case, nucleic acid, quarantine, China, life, and confirm	
13	Work resumption, lift the lockdown, neighborhood, mask, hero, test, COVID-19, nucleic acid, go out, and life	

Table 3 indicates that the topics mined by the TOT model mainly include the current status of hospital and community infections (topics 0 and 6), discussion of virus type at the beginning of the outbreak (topic 5), attention to the situation of hospital patients during the severe pandemic period (topic 7), discussion on work resumption in the later period of the pandemic (topics 2, 8, 10, and 13), and attention to the mourning of heroes and the day of lifting the lockdown. The difference in topic-word distribution between these topics obtained by the TOT model were less varied, and the same words appeared more frequently on different topics. Among the figures in the last column of Table 3, the red lines are fitted results of the beta distribution of each topic over time, while the blue lines are trends of each topic's popularity over time. The abscissa represents the date from December 2019 to April 2020 and is labelled every two weeks. However, the time distribution of each topic was relatively concentrated, had obvious peaks most of the time, and was quite independent between different topics.

In contrast to the TOT model, topics mined by the LDA model mainly include diagnosis and treatment of hospital patients (topic 3), saluting of the medical staff (topic 4), relatives and family care during the Lunar New Year (topic 5), community living materials (topic 10), attention to virus information (topic 11), and confirmed cases and work resumption (topic 12). From the results yielded by LDA model, there was a difference in topic-word distribution among the topics, and generally dissimilar words appeared in different topics. The figures in Table 4 show that the popularity of most topics is evenly distributed throughout the study period, compared to the results of the TOT model.

Table 4. Topic-clustering results of the latent Dirichlet allocation (LDA) topic model. (The blue lines are trends of each topic's popularity over time).

Topic ID	Words Describing the Topics of the LDA Model	Popularity of Each Topic over Time
3	Hospital, patient, quarantine, test, nucleic acid, doctor, community, therapy, examine, and CT	
4	Frontline, appreciate, medical workers, anti-epidemic, people, hero, China, support, national, and fight	
5	Lockdown, quarantine, friend, at home, celebrate the spring festival, family, city, message, people, and government	
10	Neighborhood, community, volunteer, resident, estate, groupon, supply, staff, worker, and proprietor	
11	Coronavirus, pneumonia, novel, virus, infect, diary, COVID-19, asymptomatic, infected person, and article	
12	Work resumption, new, case, confirm, firm, lift the lockdown, prevent and control, work, on duty, and proof	

Tables 3 and 4 indicate a considerable difference between topics generated by the LDA and TOT models. The LDA model strengthens the topic-word distribution differentiation among different topics, which means weakening the frequency of using the same words to describe different topics, whereas the TOT model emphasizes their time distribution differentiation, which means that different topics have different trends over time. The TOT model assumes that one document in the corpus consists of multiple categories of topics in some proportion, each topic has its corresponding variational trend over time within the study period, and the topic-time distribution of different topics can be as discriminable as possible. Compared with the LDA model, the time distribution of each topic mined by the TOT model is more concentrated and has an obvious peak value. Because the probability density value of topic-time distribution is larger than that of topic-word distribution, the topic-clustering results are greatly affected by time. Meanwhile, the TOT model assumes that each document consists of multiple topics; thus, the added probability distribution of topic-time weakens the impact of topic-word distribution assumption for a single topic. Therefore, the words in each document can be assigned arbitrarily to different topics, with each topic clustered almost entirely according to the temporal distribution. Consequently, several same words can be discovered to describe different topics, thus leading to the small distinction of words under various topics.

Above all, the TOT model adds the probability distribution to the model assumption, which represents the influence of temporal information on the topic-clustering results. However, each document consists of multiple topics in the TOT model, thus contributing little to short text topic clustering.

4.3. Spatiotemporal Analysis and Discussion on the Mining Results for the DMMOT Model

The DMMOT model is a time-topic model for short text topic clustering that this study built. The DMMOT model assumes that one document belongs to only one topic, and that the creation time of the document also affects topic clustering. We conducted experiments based on a microblog corpus using the DMMOT model. Table 5 shows the word distribution of all topics when the number of topics was set to 15. From the detailed description of topics generated from the DMMOT model, the hot topics of public attention focus on work and family, and the hope for an early victory against the epidemic. While the characteristics of the virus's human-to-human spread and an enthusiastic discussion of the infection sources are getting very little public attention.

Regarding the topic-word distribution, our proposed model mined many topics. Compared with the DMM model, although the same words still appeared several times in topic-words of different topics, the documents created simultaneously have greater probability of belonging to the same topic because of the addition of the time variable in the model. Thus, the influence on topic sampling probability of the co-occurrence words in one document is weakened.

Compared with the TOT model, the DMMOT model assumes that one document belongs to only one topic, rather than a combination of multiple topics, and it conforms to the real situation in which one document contains only one timestamp. The DMMOT model thus weakens the independence of different topic trends over time, which means that while some overlap areas could be obtained between different topic-time distributions, each topic-time distribution remains as separate as possible.

When the LDA model is employed for topic clustering, one document consists of multiple topics. There are still some problems in analyzing topic evolution by simply attributing the document to a certain topic according to the maximum value of the document-topic probability and establishing the relationship between the document's topic and its spatiotemporal label according to its topic category. Our proposed DMMOT model improves the limitations of the DMM and TOT models and weakens the phenomenon in which the same words appear frequently in different topics in the DMM model by cooperating with temporal information; the temporal influence on topic generation in the TOT model is also weakened by adding the assumption that a single document belongs to one topic.

When adopting our model for topic clustering, a document's topic category can be obtained directly. Accordingly, temporal distribution of a topic can be obtained by connecting the document's topic and temporal attributes. Therefore, the DMMOT model can better mine topic information of microblogs, and it can be further utilized for spatiotemporal analysis of public opinion topics in microblogs.

Table 5. Topic-clustering results of the DMMOT topic model.

Topic ID	Topic Summary	Words Describing the Topics of the DMMOT Model	Number of Documents
0	Expect and Pray	Lockdown, city, life, spring, safe, early, day, sakura, expect, and work resumption	6156
1	Material Donations	Hospital, supply, mask, patient, donate, pneumonia, medical workers, frontline, protective clothing, and Huoshenshan	999
2	Infection and Patients	Hospital, patient, quarantine, community, confirm, nucleic acid, mom, doctor, infect, and pneumonia	1583
3	Work and Family	Work resumption, mask, mom, at home, go out, life, lockdown, on duty, work, and dad	9922
4	Global Pandemic	China, country, virus, COVID-19, people, world, life, global, quarantine, and pneumonia	4116
5	Care for Family and Friends	Mask, go out, lockdown, at home, hospital, quarantine, friend, family, safe, and pneumonia	5521
6	Virus Profile	COVID-19, virus, pneumonia, patient, infect, coronavirus, China, America, test, and novel	1771
7	Community Epidemic	Community, neighborhood, patient, quarantine, confirm, resident, new, case, citywide, and prevent and control	299
8	Policy of Work Resumption	work resumption, test, neighborhood, nucleic acid, mask, go out, lift the lockdown, on duty, quarantine, and firm	4316
9	Appreciation and Salutation	Appreciate, anti-epidemic, hero, people, frontline, China, salute, hospital, medical workers, and national	3361
10	Epidemic Report	New, case, confirm, death, cumulative, suspected case, number, suspect, data, and discharge	784
11	Medical Work	Patient, hospital, protective clothing, work, mask, quarantine, on duty, nurse, and frontline, and medical workers	2730
12	Community Supply	Neighborhood, community, mask, go out, supermarket, groupon, at home, volunteer, confirm, and estate	3385
13	Work Influence	Work resumption, enterprise, work, prevent and control, firm, influence, life, pneumonia, COVID-19, and situation	1696
14	Virus Mechanism	Virus, America, Trump, mechanism, rise, person to person, China, society, COVID-19, and at present	135

4.3.1. Trends of Public Opinion Topics over Time

To measure the popularity of topics by the number of microblogs, we analyzed the temporal and spatial distribution characteristics of the popularity of different topics. Figure 3 shows the popular trends in topics over time. The blue curve indicates how the number of documents changes over time for different topics. The red curve is the result of the topic-time posterior probability distribution of the DMMOT model, that is, the fitted curve with a beta distribution.

From the results mined using the DMMOT model, the topics were relatively concentrated over time. For example, topic 5 (Care for Family and Friends) emphasized mutual care of family members and friends, so it mainly focused on the beginning of lockdown measures (23 January) and the Chinese New Year (25 January). Topic 8 (Policy of Work Resumption) was about the resumption of work after lifting the lockdown, with public attention focused on prevention and control measures such as nucleic acid testing and quarantine before returning to work. Topic 9 (Appreciation and Salutation) peaked on April 4, the national day of mourning, and its content was about paying tribute to medical workers to show appreciation to heroes of the pandemic. During the severe period of the pandemic, the public was mainly concerned with the situation of daily necessities during the lockdown, which was consistent with topic 12 (Community Supply). As the pandemic's severity gradually eased, the public began to worry and increased their concern about its impact on work and daily life, which is consistent with the content of topic 13 (Work Influence). However, some topics were widely distributed over time and did not concentrate within a certain period. For example, the popularity of topic 6 (Virus Profile) continued throughout the study period since the outbreak of the pandemic.

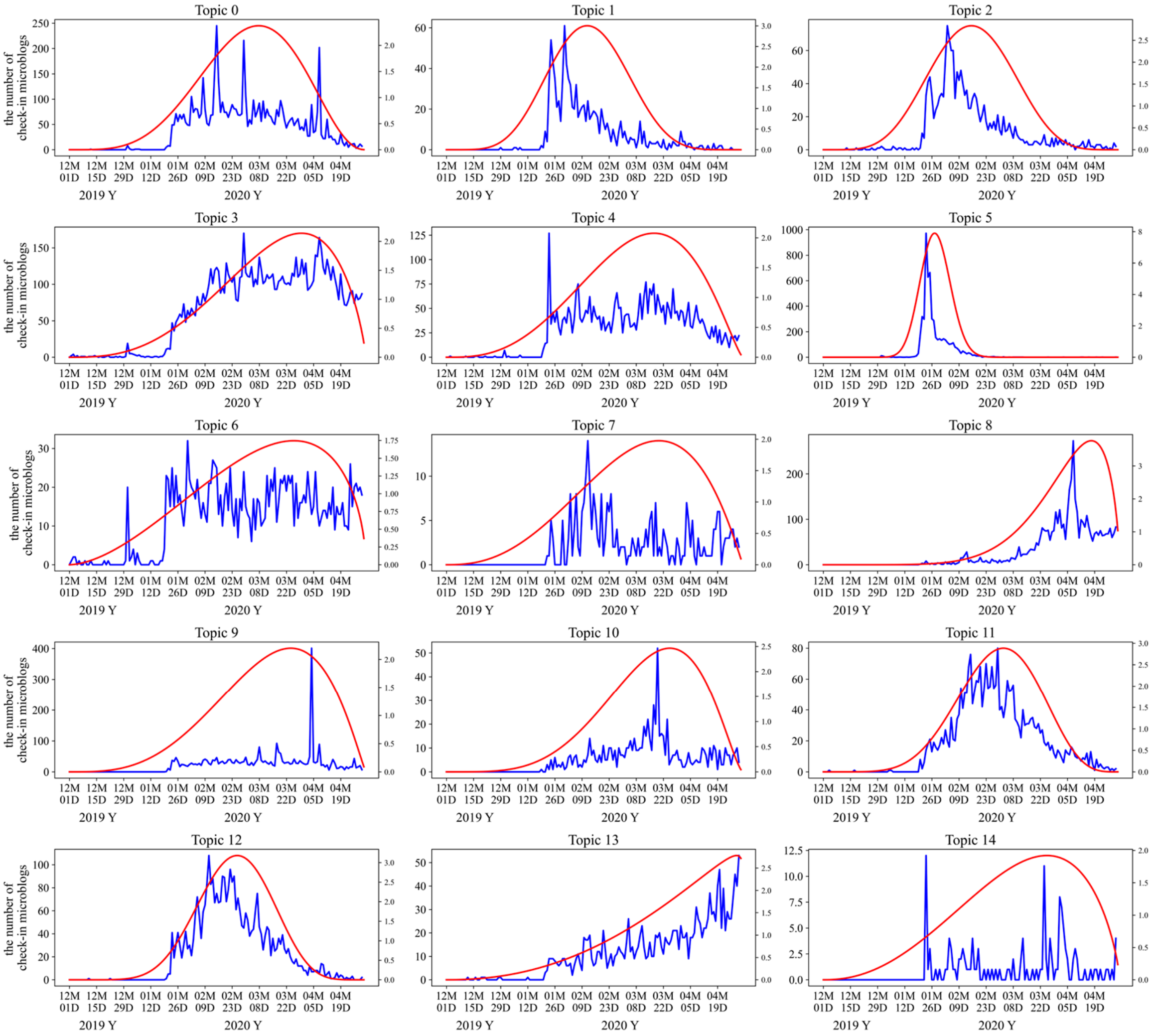
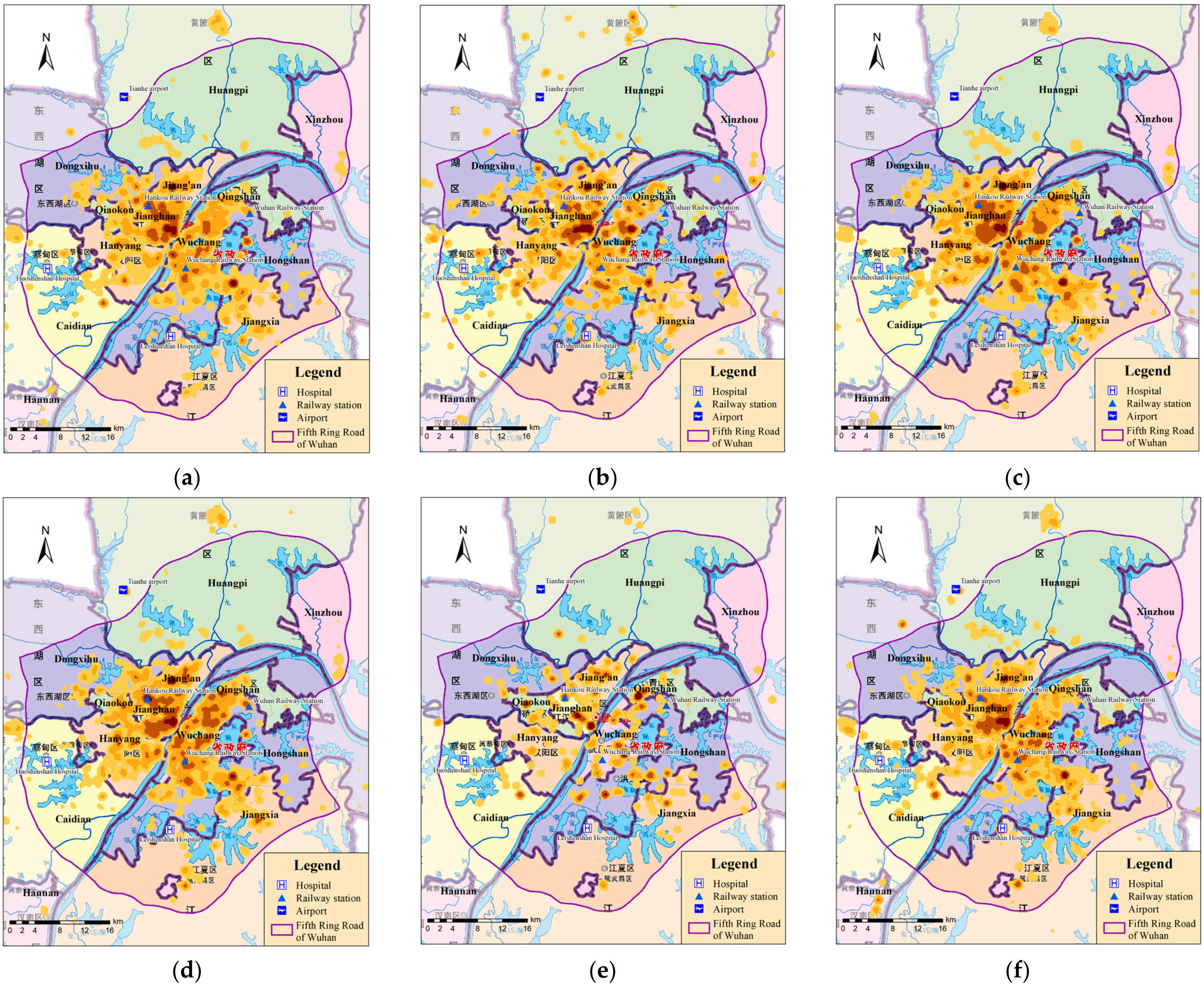


Figure 3. Variation trend of the number of check-in microblogs under different topics.

Most trends in the popularity of topics were consistent with the topics' content, and the topics peaked on festivals or days when important events occurred. The changing trend of the time series under different topics was closely related to the topic content, verifying that the text clustering results of the DMMOT model were effective.

4.3.2. Spatial Distribution of Public Opinion Topics

As hotspots of check-in microblogs are closely related to population density, and Wuhan City has high population density in the city center, the spatial distribution of public opinion topics emerges as a phenomenon of aggregation. Most check-in microblogs were concentrated within the Fifth Ring Road of Wuhan, and hotspots outside the Fifth Ring Road were mainly in the residential areas of the suburban district and Tianhe Airport. In this context, this study focused on analyzing the distribution of public opinion on the Fifth Ring Road of Wuhan. Maps of Wuhan City with administrative areas were obtained from the map world website. Figure 4 shows the kernel density distribution map of topics within and around the Fifth Ring Road of Wuhan, and deeper the color it presents, higher is the density of check-in microblogs.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 4. Kernel density estimation results of COVID-19-related check-in microblogs under different topic categories. (**a**) Topic 0; (**b**) Topic 1; (**c**) Topic 3; (**d**) Topic 5; (**e**)Topic 7; (**f**) Topic 9. All results were calculated with a bandwidth of 1000 m and cell size of 200 m × 200 m. The natural breaks (Jenks) method was used for hierarchical display.

Spatial distributions of most topics were similar to that of topic 0 (Expect and Pray), which was concentrated in the residential areas of Wuhan Central District. As shown in Figure 4a, the hotspots of topic 0 (Expect and Pray), concerning extensive residents' discussions, were mainly concentrated in Jianghan District and other areas with large residential populations. The spatial distribution of topic 1 (Material Donations) in Figure 4b was wide-ranging and maintained high popularity in the central area of Wuhan. Some hotspots were distributed around Leishenshan Hospital and Huoshenshan Hospital. This kind of distribution indicated that users discussed donations in almost every municipal district, and the phenomenon of donating medical supplies appeared in most parts of Wuhan.

Topics 3 (Work and Family) and 13 (Work Influence) were related to resumption of work. Most topics about resumption of work appeared after the lockdown was lifted, so the scope of public action had widened. Therefore, regarding the spatial distribution of the hotspots, they were more evenly distributed in the living areas on both sides of the Yangtze River. Topic 5 (Care for Family and Friends) is similar to topic 3 (Work and Family), and they were both popular and more widely distributed; topic 5 had significantly more hotspots near the three railway stations compared with other topics because public mobility was strengthened around the lockdown. Topic 7 (Community Epidemic) involved the epidemic in the community, so the hotspots were scattered. The hotspots of topic 9 (Appreciation and Salutation) appeared near Leishenshan and Huoshenshan hospitals in Wuhan, and the description of seeing off medical staff at the airport also led to the emergence of public opinion hotspots near Tianhe Airport.

Thus, based on the public opinion analysis of microblog data during the pandemic, we mined several topics, most of which are related to population density distribution and mainly concentrated in the central city of Wuhan. However, topics related to the diagnosis and treatment of hospital patients, relief supplies, gratitude, and tributes appeared as obvious hotspots in the vicinity of Leishenshan and Huoshenshan hospitals, while the airport and railway stations were also part of the gathering areas of public opinion.

5. Conclusions

The analysis of online public opinion under the pandemic situation is of great significance to understand the issues of public concern as well as grasp and guide the developing trend of public opinion over time. Based on the assumption that a single document belongs to a single topic, our study added the influence of temporal information to the topic of the document and proposed a DMMOT model. This proposed topic model analyzed the evolution of online public opinion topics based on check-in microblog data from Wuhan from 1 December 2019, to 30 April 2020. We further analyzed the temporal and spatial distribution characteristics of the topic hotspots. The following conclusions were drawn:

(1) From the perspective of the model-generation process, the assumption of DMMOT model about the document's topic made it possible to obtain a document's topic directly from the assigned results of Gibbs sampling. Furthermore, we could get the fitted topic-time distribution and combine the spatial information with the topic for spatiotemporal analysis of public opinion topics.

(2) The proposed DMMOT model performed better than the LDA, DMM, and TOT models for public opinion topic mining based on microblog data. The mining results indicated that topic-word distribution among different topics generated by the DMMOT model is differentiated, and the topic-word distribution within various topics is semantically aggregated. Meanwhile, the microblog text under each topic was gathered in a certain time window because of the topic-time distribution in the model assumption.

(3) From the perspective of the temporal and spatial distribution of public opinion topics, the topic-time distribution, obtained by the DMMOT model, generated topics that were relatively concentrated in the time window, and the characteristics of the trends of various topics over time were basically consistent with the corresponding topic content. Spatial distributions of all topics were concentrated in residential areas, and detailed distribution of the hotspots was related to the summaries of topics. Further, spatial distribution of different public opinion topics can help identify hotspots of public opinion distribution, perform differentiated public opinion management, and guide public opinion accurately.

However, this approach has certain limitations in its present state, which mark the direction of our future research. First, evaluation of the topic model's effectiveness is left for future research. Second, the beta distribution hypothesis of topic trends over time does not completely fit all topics. Therefore, future research will involve further analysis of the distribution characteristics of topic trends over time.

Author Contributions: Conceptualization, Qin Liang and Chunchun Hu; methodology, Qin Liang, Chunchun Hu, and Nianxue Luo; software, Qin Liang; validation, Qin Liang and Chunchun Hu; formal analysis, Qin Liang and Shuixiang Lu; investigation, Qin Liang and Shuixiang Lu; resources, Qin Liang and Nianxue Luo; data curation, Qin Liang; writing—original draft preparation, Qin Liang and Chunchun Hu; writing—review and editing, Qin Liang, Chunchun Hu, Nianxue Luo, and Shuixiang Lu; visualization, Qin Liang; supervision, Chunchun Hu and Nianxue Luo; project administration, Chunchun Hu and Nianxue Luo; funding acquisition, Nianxue Luo and Chunchun Hu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under grant number 2022YFC3105103, and the Natural Science Foundation of Hubei Province of China under grant number 2022CFB194.

Data Availability Statement: The data that support the findings of this study are available at https://github.com/nghuyong/weibo-cov (accessed on 24 June 2020).

Acknowledgments: The authors would like to thank Yong Hu, Heyan Huang, Anfan Chen, Xian-Ling Mao for providing supporting data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y. Revisiting several basic geographical concepts: A social sensing perspective. *Acta Geogr. Sin.* **2016**, *71*, 564–575.
2. Yang, W.; Mu, L.; Shen, Y. Effect of climate and seasonality on depressed mood among twitter users. *Appl. Geogr.* **2015**, *63*, 184–191. [CrossRef]
3. Bird, D.K.; Haynes, K.; van den Honert, R.; McAneney, J.; Poortinga, W. Nuclear power in Australia: A comparative analysis of public opinion regarding climate change and the Fukushima disaster. *Energy Policy* **2014**, *65*, 644–653. [CrossRef]
4. Shibuya, Y.; Tanaka, H. Public sentiment and demand for used cars after a large-scale disaster: Social media sentiment analysis with facebook pages. *arXiv* **2018**, arXiv:1801.07004.
5. Karami, A.; Shah, V.; Vaezi, R.; Bansal, A. Twitter speaks: A case of national disaster situational awareness. *J. Inf. Sci.* **2020**, *46*, 313–324. [CrossRef]
6. El Barachi, M.; AlKhatib, M.; Mathew, S.; Oroumchian, F. A Novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *J. Clean. Prod.* **2021**, *312*, 127820. [CrossRef]
7. Belcastro, L.; Cantini, R.; Marozzo, F. Knowledge discovery from large amounts of social media data. *Appl. Sci.* **2022**, *12*, 1209. [CrossRef]
8. Jiang, Y.; Liang, R.; Zhang, J.; Sun, J.; Liu, Y.; Qian, Y. Network public opinion detection during the coronavirus pandemic: A short-text relational topic model. *ACM Trans. Knowl. Discov. Data* **2022**, *16*, 52. [CrossRef]
9. Sina Weibo Data Center. Weibo User Development Report in 2020. Available online: https://data.weibo.com/report/reportDetail?id=456 (accessed on 16 March 2021). (In Chinese).
10. Ye, X.; Li, S.; Yang, X.; Qin, C. Use of social media for the detection and analysis of infectious diseases in China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 156. [CrossRef]
11. Yin, J.; Wang, J. A Dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014.
12. Wang, X.; McCallum, A. Topics over time: A non-markov continuous-time model of topicassl trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006.
13. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999.
14. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
15. Blei, D.M.; Lafferty, J.D. Correlated Topic Models. In Proceedings of the 19th Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005.
16. Li, W.; Sun, L.; Zhang, D. Text classification based on labeled-LDA model. *Chin. J. Comput.* **2008**, *31*, 620–627. [CrossRef]
17. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013.
18. Ma, T.; Li, J.; Liang, X.; Tian, Y.; Al-Dhelaan, A.; Al-Dhelaan, M. A time-series based aggregation scheme for topic detection in Weibo short texts. *Phys. A Stat. Mech. Its Appl.* **2019**, *536*, 120972. [CrossRef]
19. Walde, S.S.I.; Melinger, A. An in-depth look into the co-occurrence distribution of semantic associates. *Ital. J. Linguist.* **2008**, *20*, 89–128.
20. Li, C.; Duan, Y.; Wang, H.; Zhang, Z.; Sun, A.; Ma, Z. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Trans. Inf. Syst.* **2017**, *36*, 1–30. [CrossRef]
21. Rahimi, M.; Zahedi, M.; Mashayekhi, H. A probabilistic topic model based on short distance co-occurrences. *Expert Syst. Appl.* **2022**, *193*, 116518. [CrossRef]

22. Blei, D.M.; Lafferty, J.D. Dynamic Topic Models. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006.
23. Han, X.; Wang, J.; Zhang, M.; Wang, X. Using social media to mine and analyze public opinion related to COVID-19 in China. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2788. [CrossRef] [PubMed]
24. Wang, J.; Zhang, M.; Han, X.; Wang, X.; Zheng, L. Spatio-temporal evolution and regional differences of the public opinion on the prevention and control of COVID-19 epidemic in China. *Acta Geogr. Sin.* **2020**, *75*, 2490–2504.
25. Boon-Itt, S.; Skunkan, Y. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill.* **2020**, *6*, 245–261. [CrossRef]
26. Amara, A.; Hadj Taieb, M.A.; Ben Aouicha, M. Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Appl. Intell.* **2021**, *51*, 3052–3073. [CrossRef] [PubMed]
27. Hu, Y.; Huang, H.; Chen, A.; Mao, X.L. Weibo-COV: A large-Scale COVID-19 social media mataset from Weibo. *arXiv* **2020**, arXiv:2005.09174.
28. Hu, Y.; Huang, H.; Chen, A.; Mao, X.L. Weibo-COV 2.0. 2020. Available online: https://github.Com/nghuyong/weibo-cov (accessed on 24 June 2020).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.