Research on Local-Global Spatio-Temporal Topic Model based on Social Media Text Data with Location Information

Haiqi Wang [1], Xueying Li [1], Fadong Li [1], Yawen Ou [1], Yuanhao Cao [1], Baozhong Wang [1], Jun He [1], Tong Liu [1], Yanwei Wang [1]

[1] China University of Petroleum, College of Oceanography and Space Informatics, Qingdao, China – wanghaiqi@upc.edu.cn, z23160027@s.upc.edu.cn, z21160098@s.upc.edu.cn, s23160010@s.upc.edu.cn, z23160031@s.upc.edu.cn, baozhongw@foxmail.com, z24160039@s.upc.edu.cn, z24160034@s.upc.edu.cn, 20240064@upc.edu.cn.

Keywords: Spatio-temporal topic modeling, Biterm Topic Model, Short text analysis, Local-global topics, Social media mining, Geographic information retrieval.

Abstract

Social media check-in data contains textual, temporal and spatial information, which is of great value for extracting topics and analyzing spatio-temporal changes to capture online public opinion trends. However, because existing topic models rely on predefined functions or distributions, they struggle to handle short text data sets and separate local and global topics. In this study, a Spatio-Temporal Biterm Topic Model (ST-BTM) is proposed, which integrates word pair modeling, spatio-temporal slicing and local-global topic extraction framework. ST-BTM uses spatio-temporal information from social media data to extract global topics while identifying local topics closely related to a specific region in local documents, thereby analyzing the spatio-temporal changes of topics in detail. Experiments on the Weibo COVID-19 topic check-in dataset show that ST-BTM model's UMass average consistency scores under different topic numbers are 50%, 19% and 8% higher than LDA, BTM and ToT models, respectively. Experiment show that this method captures cross-region local topics effectively, proves its ability to process short text data sets

1. Introduction

During large-scale events such as the COVID-19 pandemic, user-generated content on platforms like Weibo becomes a rich source of real-time spatial and temporal public opinion data. In the field of natural language processing (NLP), topic models are widely used tools for analyzing textual data and have been extensively applied in COVID-19-related studies (Liang, 2021; Gangwar and Singh, 2021; Zhang, 2021; Ghasiya and Okamura, 2021; Garcia and Berton, 2021).

Spatio-temporal topic models are more complex and comprehensive as they incorporate both temporal and geographic information of the text (Mei, 2006). Liu proposed upstream and downstream spatio-temporal topic models, considering the varied effects of time on topics. Using these models to analyze user check-in data, they were able to extract spatio-temporal information on user mobility and interests (Liu, 2015). He introduced a probabilistic latent generative model, the spatio-temporal topic model, designed to infer personal interests, the spatio-temporal patterns of topics embedded in user check-in activities, the interdependencies between category-topic and sentiment-topic, as well as the correlations between sentiment labels and rating scores (He, 2017). Li proposed a Spatio-Temporal Topic Model (STTM) for cold-start event recommendation, which captures users' evolving interests in content and geographic spaces over time. In STTM, users exhibit varying distributions of event and location topics at different time periods (Li, 2020).

Topic models primarily focus on extracting overall thematic information from documents, referred to as Global Topics. However, Local Topics are frequently overlooked during the analysis process. These topics respectively describe the thematic structure of the dataset at different hierarchical levels. Qiang distinguished local and global topics by filtering out region-irrelevant words using varying weights and leveraging topic generation probabilities (Qiang, 2017). Liu trained Global Topics using Weibo data from all regions and learned Local Topics for a specific location by introducing a Bernoulli distribution in their model to differentiate whether a document belongs to a local or global context (Liu, 2018).

In the aforementioned studies, various spatio-temporal topic models have been constructed to address different research objectives and subjects. However, traditional topic models have been extensively used to extract thematic structures from large-scale textual corpora, they often overlook or underperform in modeling short texts with fine-grained spatio-temporal characteristics. Moreover, many topic models conflate local topics—those specific to a region—with global topics that span multiple regions, thereby failing to capture regional heterogeneity in public discourse.

This study focuses on social media text data with location and constructs a spatio-temporal topic model (ST-BTM) tailored to the textual characteristics and spatio-temporal attributes of the dataset, enabling the simultaneous extraction of topics and the analysis of their spatio-temporal dynamics. Building on this foundation, an innovative Local Topics extraction framework is proposed, ensuring the identification of Global Topics while capturing region-specific Local Topics.

2. Methods

2.1 ST-BTM Spatio-temporal Topic Model

2.1.1 Model Construction: This study uses the administrative city level as the smallest spatial unit for segmentation, ensuring the independence of data from different regions. Based on this, the document data from different regions is further segmented by day as the smallest time unit. This results in multiple spatiotemporal document sets, as shown in Figure 1.

Each document in the dataset can be represented as $d(l,t,w)$, where l indicates the region the document belongs to, t is the document's publication time, and w denotes the words contained in the document. After dividing the documents by region and date, multiple spatio-temporal document sets $s(l,t)$ are formed. For each spatio-temporal document set s, l represents the region

associated with the document set, and t corresponds to the specific date of the document set.

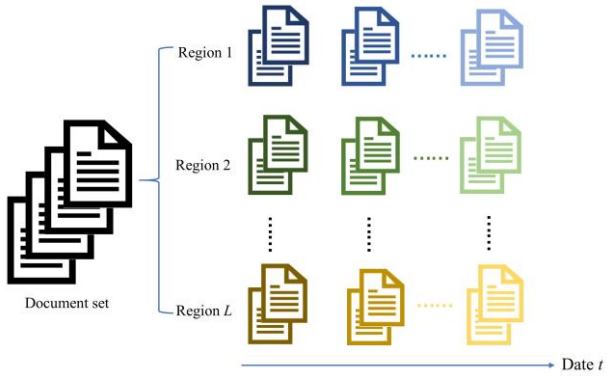

Figure 1. Schematic diagram of spatio-temporal document collection partitioning

Based on the BTM model, this study integrates the temporal information processing method from the ToT model into the BTM framework. Additionally, considering the characteristics of the data after segmentation, a spatiotemporal topic model, named ST-BTM, is constructed. The graphical structure of the ST-BTM model is shown in Figure 2.

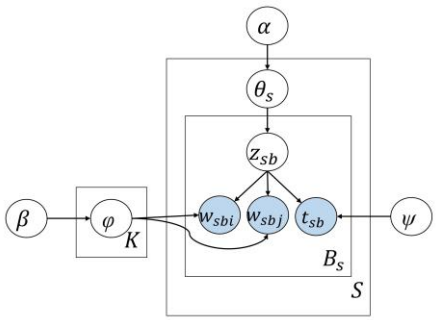

Figure 2. Graphic model representation of the ST-BTM.

The ST-BTM model not only captures more accurate topic information but also provides insights into the temporal evolution of topics and the distribution of topics across different regions. In the model, each spatio-temporal slice corresponds to a unique set of biterms B_s and the topic distribution θ_s for that spatio-temporal slice. The biterms generated by each document are determined by the words within that document. Each biterm can be represented as (w_{sbi},w_{sbj}), where w_{sbi} represents the i-th word of the b-th biterm within spatio-temporal slice s. Words from each document are paired in an unordered manner to form the biterms set associated with that document.

In the model, β and α are the parameters of the dirichlet prior distributions for φ and θ_s, respectively. Both φ and θ_s follow a multinomial distribution, where φ represents the distribution of words in each topic and θ_s represents the topic distribution in each spatio-temporal slice. In addition, z_{sb} denotes the topic assignment for the biterm representing the b-th pair of words in slice s. The two words in each biterm belong to the same topic. In the model, all words belonging to the same spatio-temporal slice share the same time attribute. Here is the specific generative process for ST-BTM:

(1) For each topic $z \in \{1,...,K\}$, obtain the multinomial distribution φ_z from Dir(β);

(2) For each space-time slice $s \in \{1,\ldots,S\}$;

a. Obtain the multinomial distribution θ_s from Dir(α);

b. For each Biterm $b \in \{1,\ldots,B_s\}$

i. Obtain z_{sb} from Mult(θ_s);

ii. Obtain the words w_{sbi} and w_{sbj} from Mult(φ_z);

iii. Obtain t_{sb} from Beta(ψ_z).

2.1.2 Derivation of the Joint Distribution Formula for the ST-BTM Model: By decomposing the graphical structure of the Spatio-Temporal Biterm Topic Model (ST-BTM), the first step is to obtain the topic distributions of biterm sets generated by documents across different spatio-temporal regions, as illustrated in Figure 3.

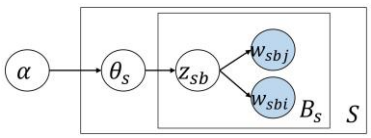

Figure 3. Process for acquiring topics from biterms collection.

For a specific spatio-temporal slice s, assume that all biterms and their corresponding topics within this slice are represented by Equations (1) and (2).

$$b_s = \left(b_1, b_2, b_3, \ldots, b_{|Bs|}\right) \tag{1}$$

$$z_s = \left(z_1, z_2, z_3, \ldots, z_{|Bs|}\right) \tag{2}$$

$$p(z \mid \alpha) = \prod_{s=1}^{S} p(z_s \mid \alpha) = \prod_{s=1}^{S} \frac{\Delta(n_{sz} + \alpha)}{\Delta(\alpha)} = \prod_{s=1}^{S} \frac{\frac{\prod_{z=1}^{K} \Gamma(n_{sz} + \alpha_z)}{\Gamma(\sum_{z=1}^{K}(n_{sz} + \alpha_z))}}{\frac{\prod_{z=1}^{K} \alpha_z}{\Gamma(\sum_{z=1}^{K}(\alpha_z))}} \tag{3}$$

Let b_s denote the set of all biterms within the spatio-temporal slice, where a biterm is defined as $b_1=(w_{1i}, w_{1j})$. z_s represents the topic assignments corresponding to each biterm in b_s, with $z_1=(z_{1i}, z_{1j})$, and it holds that $z_{1i}=z_{1j}$. As shown in the Figure 3, the process $\alpha \rightarrow \theta_s \rightarrow z_{sb}$ represents the generation of topics corresponding to all words in the biterm set. Here, θ_s follows a Dirichlet distribution with parameter α, and z_{sb} follows a multinomial distribution parameterized by θ_s. This forms a conjugate structure between the Dirichlet and multinomial distributions. Based on this conjugate structure and the theory of the Dirichlet distribution, the posterior distribution of the parameter θ_s is given by: $Dir(\theta_s|n_{sz}+\alpha)$, where n_{sz} represents the set of counts of words generated by different topics z within spatio-temporal slice s. Given the posterior distribution of θ_s, the topic assignments for the biterm set within this spatio-temporal slice can be computed.The topic generation processes for different spatio-temporal slices are independent of each other. Consequently, the topic generation process for all biterms in the entire corpus can be regarded as S independent Dirichlet-multinomial conjugate structures. The corresponding probability distribution is given in Equation (3). Here, z represents the set of topic assignments for biterms across different spatio-temporal slices.

After obtaining the topic assignments for the biterm sets in each spatio-temporal slice, it is necessary to further determine the

word distributions within each topic, thereby deriving the topic-specific words. This process results in the generation of topic words, as illustrated in Figure 4.

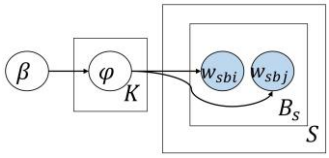

Figure 4. Process for acquiring topic words.

The topic words are calculated based on the topic assignments of all words in the corpus. Therefore, this process requires combining all biterms for unified processing. Specifically, words from the biterm set that share the same topic are grouped together, as shown in Equations (4), (5), and (6).

$$w_i^{'} = (w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ik}) \tag{4}$$

$$w_j^{'} = (w_{j1}, w_{j2}, w_{j3}, \ldots, w_{jk}) \tag{5}$$

$$z' = (z_1, z_2, z_3, \ldots, z_k) \tag{6}$$

Since a biterm consists of two words (w_i,w_j), it is divided into two groups. Let w_i' denote the set of words from the first word of the biterm, where w_{ik} represents the set of words in w_i' with topic k. Similarly, w_j' represents the set of words from the second word of the biterm, and w_{jk} represents the set of words in w_j' with topic k. w_i' and w_j' are essentially identical but are differentiated for convenience in the subsequent calculations. In z', z_k represents the topic corresponding to the sets w_i' and w_j', which include w_{ik} and w_{jk}, respectively. As shown in the Figure 4, the process of generating all words for topic k follows the structure $\beta \rightarrow \varphi_k \rightarrow w_{ik}|k=z$ and $\beta \rightarrow \varphi_k \rightarrow w_{jk}|k=z$, where k is determined during the process of generating topics for the spatio-temporal biterm sets. Here, $\beta \rightarrow \varphi_k$ represents a Dirichlet distribution with parameter β, and $\varphi_k \rightarrow w_{ik}$ and $\varphi_k \rightarrow w_{jk}$ are multinomial distributions. This means the topic word distribution for topic k corresponds to two conjugate structures. All topics thus correspond to $2K$ Dirichlet-multinomial conjugate structures. Consequently, the posterior distribution of φ can also be computed. By combining the Dirichlet distribution and conjugate structure theory, the generation probabilities of all words for topic k are given by Equations (7) and (8).

$$p(w_{ik} \mid \beta) = \frac{\Delta(\vec{n}_{ik} + \beta)}{\Delta(\beta)} \tag{7}$$

$$p(w_{jk} \mid \beta) = \frac{\Delta(\vec{n}_{jk} + \beta)}{\Delta(\beta)} \tag{8}$$

$$\begin{aligned} p(b \mid z, \beta) &= p(w_i \mid z, \beta)\, p(w_j \mid z, \beta) \\ &= p(w_i^{'} \mid z, \beta)\, p(w_j^{'} \mid z, \beta) \\ &= \prod_{k=1}^{K} p(w_{ik} \mid z_k, \beta) \prod_{k=1}^{K} p(w_{jk} \mid z_k, \beta) \\ &= \prod_{k=1}^{K} \frac{\Delta(\vec{n}_{ik} + \beta)}{\Delta(\beta)} \frac{\Delta(\vec{n}_{jk} + \beta)}{\Delta(\beta)} = \prod_{k=1}^{K} \left(\frac{\Delta(\vec{n}_{k} + \beta)}{\Delta(\beta)}\right)^2 \end{aligned} \tag{9}$$

In the Equations (10) and (11), $\vec{n}_{ik} = (n_{ik}^{(1)}, n_{ik}^{(2)} n_{ik}^{(3)}, \ldots, n_{ik}^{(V)})$ and $\vec{n}_{jk} = (n_{jk}^{(1)}, n_{jk}^{(2)} n_{jk}^{(3)}, \ldots, n_{jk}^{(V)})$, where $\vec{n}_{ik}$ and $\vec{n}_{jk}$ represent the counts of word t generated by the k-th topic for the words. Using this, the generation probability for all biterms can be further derived. In the equation, $\Delta(\vec{n}_{ik} + \beta) = \Delta(\vec{n}_{jk} + \beta)$, thus the above expression can be simplified, as shown in Equation (9).

In addition to the two processes described above, there is also the time parameter t corresponding to the topic, which follows a Beta distribution with parameter ψ. Specifically, $t_{sb}|\psi_{sb} \sim Beta(\psi_{sb})$. Here, t_{sb} represents the time associated with biterm b in spatio-temporal slice s, and the corresponding probability distribution function is given by Equation (10).

Combining the aforementioned processes, the joint distribution formula for ST-BTM is given by $P(b,t,z|\alpha,\beta,\psi)$. The results are integrated as shown in Equation (11).

$$P(t \mid z, \psi) = \prod_{s=1}^{S} \prod_{b=1}^{B_s} P(t_{sb} \mid \psi_{sb}) \tag{10}$$

$$\begin{aligned} &P(b, t, z \mid \alpha, \beta, \psi) = P(w_i \mid z, \beta) P(w_j \mid z, \beta) P(z \mid \alpha) P(t \mid z, \psi) \\ &= \int P(w_i \mid \varphi, z) P(w_i \mid \varphi, z) P(\varphi \mid \beta) d\varphi \int P(z \mid \theta) P(\theta \mid \alpha) d\theta P(t \mid z, \psi) \\ &= \left[\left(\frac{\Gamma\left(\prod_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \prod_{z=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(n_{zv} + \beta_v)}{\Gamma\left(\sum_{v=1}^{V} (n_{zv} + \beta_v)\right)} \right]^2 \left(\frac{\Gamma\left(\prod_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \right)^S \\ &\times \prod_{s=1}^{S} \frac{\prod_{z=1}^{K} \Gamma(n_{zv} + \alpha_z)}{\Gamma\left(\prod_{z=1}^{K} (n_{zv} + \alpha_z)\right)} \prod_{s=1}^{S} \prod_{b=1}^{B_s} P(t_{sb} \mid \psi_{z_{sb}}) \end{aligned} \tag{11}$$

In the equation, n_{zv} denotes the number of words assigned to topic z in the vocabulary V, and n_{zv} represents the number of biterms assigned to topic z in spatio-temporal slice s. The joint distribution formula consists of three parts: $P(w_i|z,\beta)P(w_j|z,\beta)$, $P(z|\alpha)$, and $P(t|z,\psi)$. Each part corresponds to a different process. The first part is derived from $P(b|z,\beta)$, which establishes the relationship between each biterm and its corresponding topic. The second part represents the topic distribution for each spatio-temporal slice s, and the final part captures the temporal evolution of the different topics.

2.1.3 Approximate Solution for the Parameters of the Joint Distribution Formula: This paper uses Gibbs sampling to infer the model parameters. In the joint distribution formula for the ST-BTM model, the topic z is an unobserved variable, so the only parameter that needs to be sampled is z. In this process, using BTM's Gibbs sampling as a reference, a random topic is initially assigned to each biterm in the corpus. At this point, a biterm b from spatio-temporal slice s is extracted, and the topic z_{sb} is assigned to this biterm. Then, the topics of all biterms, excluding the current biterm, are updated to assign a new topic that is closer to the true topic assignment for that biterm. The above process is the Gibbs sampling process.

2.2 Local Topics Acquisition Method

Local Topics are generated by the documents in a specific region, which reflect the unique concerns and discussion focus of the region. Effectively identifying the local topic documents and the global topic documents in the local document set is the key to obtaining the local topics.

For all documents in a specific region, each document participates in the overall topic discovery process of the document set. The topics extracted in this process are typically global topics. Each document in the local set has varying degrees of connection with the global topics. Some documents contain words with higher weights in the global topic results, making the document's meaning closely aligned with the global topics. In contrast, other documents contain more region-specific words or words related to regional hot events, resulting in a smaller similarity to the global topics. These documents' meanings differ significantly from the global topic results. To distinguish such documents that differ from the global topics, they can be divided based on their similarity to the global topics, as shown in Figure 5.

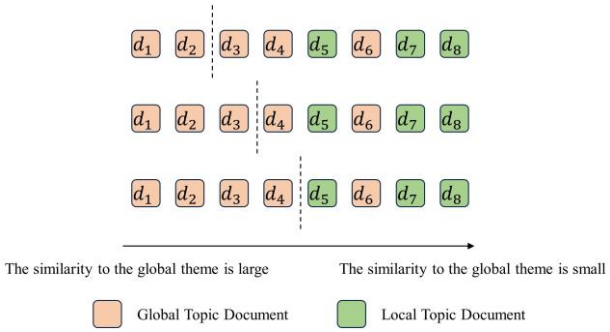

Figure 5. Document collection partitioning based on global topic similarity

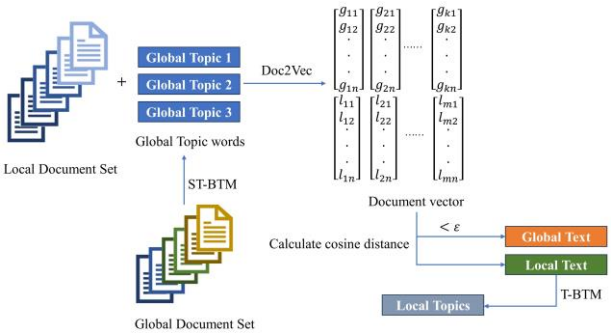

Figure 6. Local Topics acquisition process.

Documents are ranked based on their similarity to global topics, and an appropriate similarity threshold is gradually selected to partition the document set. This approach effectively distinguishes between documents with high similarity to global topics and those with low similarity, i.e., Global Topic Document (Global Text) and Local Topic Document (Local Text).

Using the above approach, a corresponding Local topic acquisition method is designed, as shown in Figure 6. The method first utilizes the ST-BTM model, constructed in this paper, to analyze the global document set in order to extract Global Topics that represent general issues of common concern across all regions. The extracted global topics are then merged with the local document set, and document vector transformation methods are applied to convert documents into corresponding document vector representations. After the document vector transformation, the cosine distance between each local document and the global topics is calculated. The cosine distance measures the similarity between the document and the global topics. By setting an appropriate threshold, local documents can be further classified into two categories: documents with a high similarity to global topics (i.e., cosine distance smaller than the threshold) are classified as global topic documents, while documents with lower similarity (i.e., cosine distance greater than or equal to the threshold) are classified as local topic documents. Finally, the short text temporal topic model T-BTM, which combines BTM model and ToT model, is used to analyze the divided regional topic documents in depth to extract accurate Local Topics

3. Experiments and Analysis

3.1 Model Performance Comparison

The dataset in this section is sourced from Weibo platform, containing check-in data with keywords such as "疫情(pandemic)", "肺炎(pneumonia)", and "新冠(COVID-19)" from December 1, 2022, to January 31, 2023. A total of 54,912 Weibo check-in records were collected. In the dataset, some regions had insufficient data, and the topic results exhibited by these sparse data were not convincing. Therefore, in the experiment, data from first-tier and new first-tier cities were further selected as the experimental data. After filtering, 24,749 Weibo records remained.

This paper uses the UMass coherence score as the quantitative evaluation metric for the topic model results. A higher value indicates stronger interpretability of the topic results. The calculation process is shown in Equation (12).

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)} \qquad (12)$$

Since the ST-BTM model is improved by the BTM and ToT models, the traditional topic models LDA, BTM and ToT models are used for model comparison.

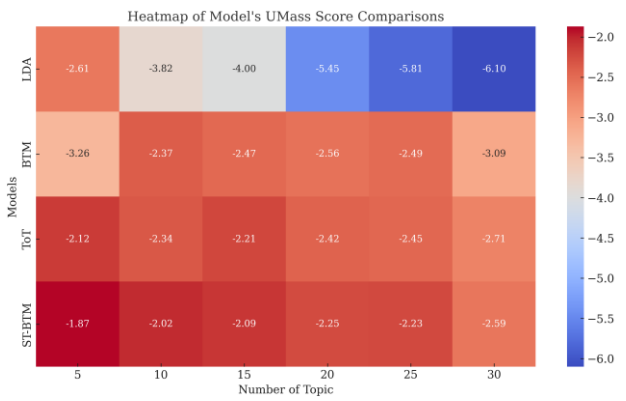

Figure 7. Comparison of UMass scores for different topic models

In the experiment, the hyperparameters for all topic models are kept consistent, where α=0.5 and β=0.1. For the " 疫情 (pandemic) " Weibo check-in dataset, as the content in the dataset mainly revolves around the pandemic topic, the

maximum number of topics is set to 30 in the experiment. Specifically, *K* values of 5, 10, 15, 20, 25, and 30 are tested. The UMass coherence scores for the topic results output by different models with varying numbers of topics are shown in Figure 7.

As shown in Figure 7, the proposed ST-BTM model consistently achieves the highest UMass scores across different numbers of topics compared to other models. The UMass scores of all four models generally decrease as the number of topics increases, with the ST-BTM model exhibiting the smallest decline. The UMass scores for different numbers of topics within the same model vary, and by calculating the average UMass score for each model across the range of 5 to 30 topics, it is observed that the ST-BTM model improves the average UMass coherence score on the COVID-19 topic check-in dataset by 50%, 19%, and 8% compared to the LDA, BTM, and ToT models, respectively. When the number of topics is set to 5, the ST-BTM model improves the UMass coherence score on the COVID-19 topic check-in dataset by 28%, 43%, and 12% compared to the LDA, BTM, and ToT models, respectively. These experimental results demonstrate that the proposed spatiotemporal topic model ST-BTM outperforms the three traditional topic models in extracting topic information.

3.2 Global Topic Extraction

The experimental data in this section is sourced from original content on the Weibo platform between January 1st and December 31st, 2022. The data focuses on Weibo posts containing the keywords "新冠(COVID-19)" and "疫情(pandemic)", and includes posts with geolocation information. A total of 265,119 geolocated Weibo posts were crawled. After data preprocessing, 164,389 valid posts were retained, while 100,730 posts were excluded as invalid.

To obtain global topics, it is necessary to determine the optimal number of topics for the dataset. In this section, the optimal number of topics is identified using the UMass coherence score. During the experiment, all model parameters, except for the number of topics *K*, are kept constant. Specifically, the model parameters α and β are set to 0.5 and 0.1, respectively, and the number of iterations Niter is set to 100. The value of *K* is varied from 1 to 50, and the corresponding UMass scores for different values of *K* are computed. A higher UMass score indicates better topic quality. The experimental results are shown in Figure 8. It can be observed that the UMass score is maximized when *K*=2. As *K* increases beyond 2, the UMass score gradually decreases, indicating that 2 is the optimal number of topics for this dataset.

Using the ST-BTM model to perform topic partitioning with *K*=2 on the dataset, the word clouds for the topics are shown in Figure 9. By summarizing the topic words of the two topics, the following two themes are identified:
(1) Personal Daily Life Impact：“希望(Hope)”, “生活(Life)”, “感觉(Feel)”, “回家(Going home)”, etc.
(2) Pandemic Control：“结束(End)”, “解封(Unblock) ”, “抗疫(Anti-epidemic)”, etc.

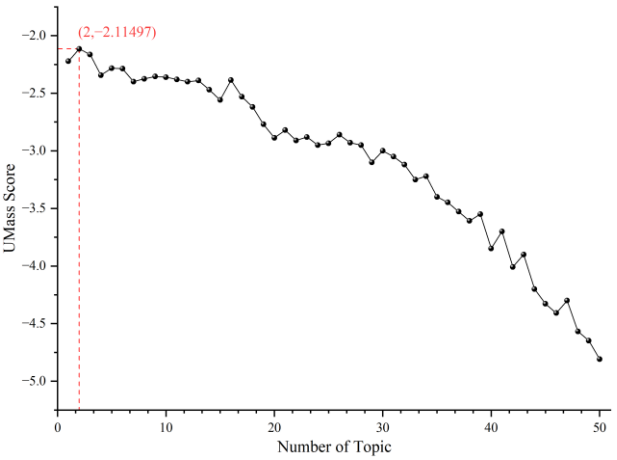

Figure 8. UMass Scores under Different Topic Numbers in COVID-19 Dataset.

Figure 9. Word cloud of Global Topic 1 (left) and Global Topic 2 (right).

3.3 Spatio-temporal difference analysis of Global Topics

Through the application of the ST-BTM model, the temporal evolution trends of Global Topic 1 (representing the pandemic’s impact on individual daily life) and Global Topic 2 (representing pandemic prevention and control measures) were also identified, as illustrated in Figure 10 and 11.

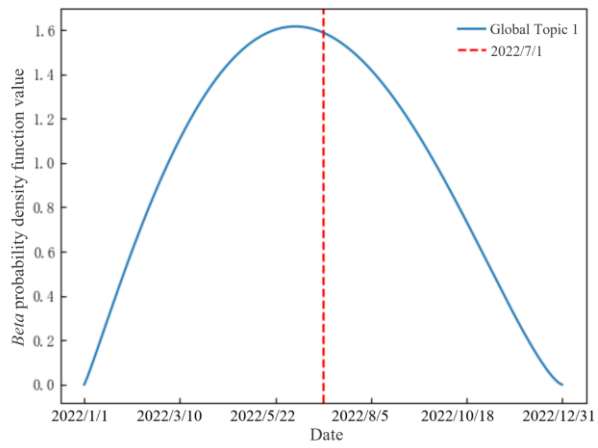

Figure 10. Time-varying Curve of Global Topics 1.

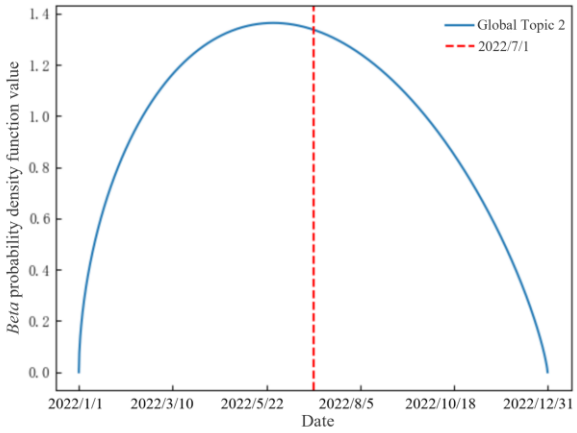

Figure 11. Time-varying Curve of Global Topics 2.

From Figure 10 and 11, it is clear that the trends of these two global topics are generally consistent, both showing a pattern of gradual increase, peaking, and then gradually declining. This indicates that public discussions on pandemic control and the impact of the pandemic on daily life followed a similar trajectory. Upon further observation, it can be noted that the variation in both Global Topic 1 and Global Topic 2 was more pronounced in the first half of the year compared to the second half. This suggests that discussions on pandemic control and its impact on daily life were more intense and widespread during the first half of the year.

To quantitatively assess the differences in the discussion intensity of different global topics across various months nationwide, the study uses the Coefficient of Variation (CV) as a quantitative indicator, as shown in Equation (13).

$$c_v = \frac{\sigma}{\mu} \tag{13}$$

In the equation, σ and μ represent the standard deviation and mean of the data, respectively. The CV provides an intuitive reflection of the uneven distribution of topic probabilities across different months. In the Figure 12, the CV for the topic probabilities of Global Topic 1 reaches its maximum in June, while the CV for Global Topic 2 peaks in November.

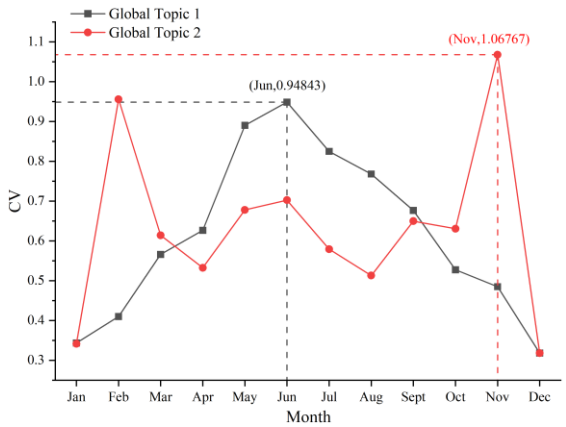

Figure 12. Monthly Coefficient of Variation for Global Topics.

For Global Topic 1, the spatial distribution of topic probabilities in June, when the CV reaches its peak, is shown in Figure 13. From the spatial distribution map of the topic probabilities for Global Topic 1 in June, it can be observed that the topic probability related to the impact of the pandemic on personal daily life is relatively low in most regions. High-probability areas are scattered and fewer in number.

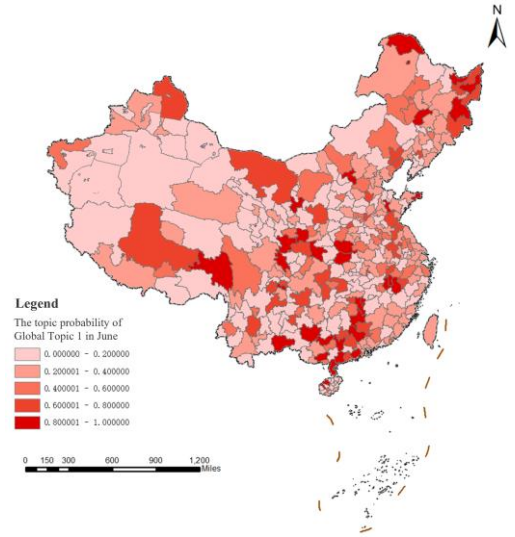

Figure 13. Spatial Distribution of Topic Probability for Global Topic 1 in June.

The spatial distribution of Global Topic 2 probabilities in November is shown in Figure 14. The map indicates that most regions had lower probabilities for this topic, concentrated in the range of 0 to 0.4, suggesting low public engagement in discussing pandemic prevention measures in these areas. However, some regions showed significantly higher probabilities.

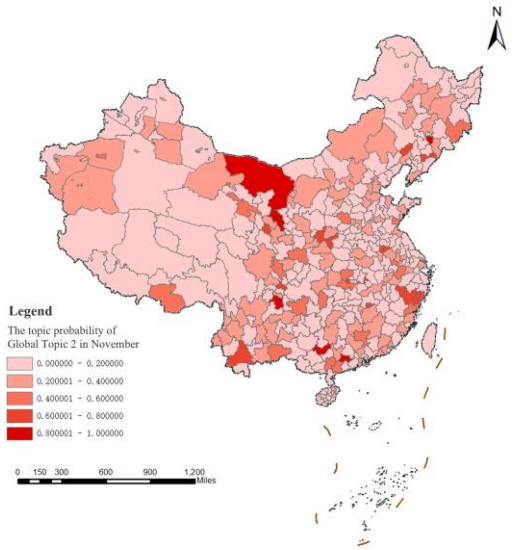

Figure 14. Spatial Distribution of Topic Probability for Global Topic 2 in November.

3.4 Local Topic Extraction

The ST-BTM model identified the optimal global topics for the dataset and subsequently extracted corresponding local topics across different regions. This section illustrates the process of

local topic extraction using Shanghai, the region with the highest volume of textual data, as a case study.

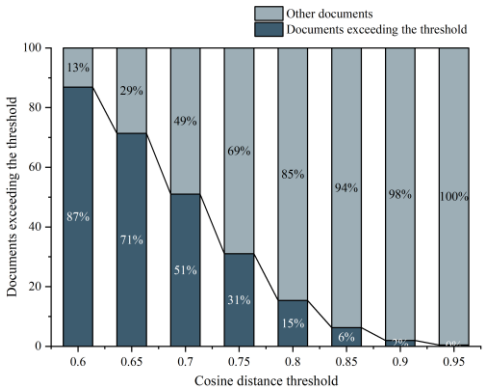

Figure 15. Division Results Under Different Cosine Distance Thresholds for Shanghai City.

First, the document set from Shanghai was integrated with the global topics. Texts were transformed into vectors using the PV-DM model, with the word embedding dimension set to 50. Cosine distances between each local document and the global topics were then computed. Based on varying cosine distance thresholds, the Shanghai document set (comprising 17,796 entries) was segmented, as shown in Figure 15. As expected, higher thresholds resulted in fewer documents meeting the inclusion criterion.

Subsequently, topic modeling is performed on each segmented document set to extract the high-frequency words. These are then compared with the top 10 words of each global topic to calculate the difference scores, as shown in Table 1. By combining the difference score with the topic coherence (UMass) score, the optimal segmentation threshold was determined to be 0.85. At this threshold, the resulting local topic exhibited substantial divergence from the global topics while maintaining a high degree of interpretability.

Using the threshold of 0.85, a local topic document set consisting of 1,119 entries was obtained. Topic modeling on this set was conducted using the T-BTM model. UMass coherence scores across varying numbers of topics shown in Figure 16, indicated that a single-topic configuration yielded the highest score, suggesting that the optimal number of topics for this subset is one.

Cosine distance threshold	Topic words (Chinese)	Topic Words (translated into English)	Topic coherence score	Difference score with Global Topic 1	Difference score with Global Topic 2
0.6	上海, 真的, 希望, 小区, 核酸, 生活, 结束, 隔离, 期间, 很多;	Shanghai, Really, Hope, Residential district, Nucleic acid, Life, End, Quarantine, Period, Plenty of	-2.37	0.75	0.75
0.65	上海, 真的, 小区, 希望, 核酸, 生活, 租房, 结束, 隔离, 期间;	Shanghai, Really, Residential district, Hope, Nucleic acid, Life, Rent a house, End, Quarantine, Period	-2.74	0.80	0.75
0.7	上海, 小区, 租房, 生活, 核酸, 期间, 物资, 结束, 隔离, 很多;	Shanghai, Residential district, Rent a house, Life, Nucleic acid, Period, Supplies, End, Quarantine, Plenty of	-3.13	0.85	0.75
0.75	上海, 租房, 小区, 生活, 核酸, 物资, 期间, 解封, 很多, 感觉;	Shanghai, Rent a house, Residential district, Life, Nucleic acid, Supplies, Period, Unblock, Plenty of, Feel	-3.84	0.90	0.75
0.8	上海, 小区, 生活, 租房, 物资, 期间, 只能, 核酸, 很多, 两个;	Shanghai, Residential district, Life, Rent a house, Supplies, Period, Can only, Nucleic acid, Plenty of, Two	-3.93	0.90	0.80
0.85	上海, 小区, 物资, 生活, 团购, 只能, 两个, 家里, 居家, 感觉;	Shanghai, Residential district, Supplies, Life, Group buying, Can only, Two, Home, At home, Feel	-2.48	0.95	0.85
0.9	上海, 家里, 居家, 小区, 几个, 一点, 鸡蛋, 生活, 发现, 感觉;	Shanghai, Home, At home, Residential district, A few, Eggs, Life, Find, Feel	-5.06	0.95	0.85
0.95	上海, 生活, 几个, 阶段, 抗原, 社会, 法餐, 居家, 少许, 生抽。	Shanghai, Life, A few, Stage, Antigen, Social, French, At home, A little, Light soy	-18.14	0.95	0.95

Table 1. Comparison of Document Topic Results Under Different Threshold Divisions and Global Topics for Shanghai City

Figure 17 reveals high-frequency terms such as “Shanghai,” “supplies,” “group buying,” and “vegetables,” highlighting local residents’ concerns regarding essential goods and their acquisition, particularly through group buying. Figure 18 shows a significant peak in early 2022, with the highest attention observed in April during a period of material shortages. As aid

from other provinces began to alleviate the situation, the intensity of discussion on this topic gradually declined.

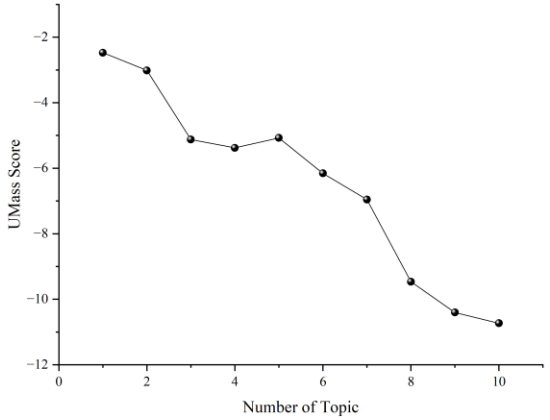

Figure 16. UMass Scores for Different Numbers of Topics After Division for Shanghai City.

Figure 17. Cloud Map of Topic Words in Shanghai

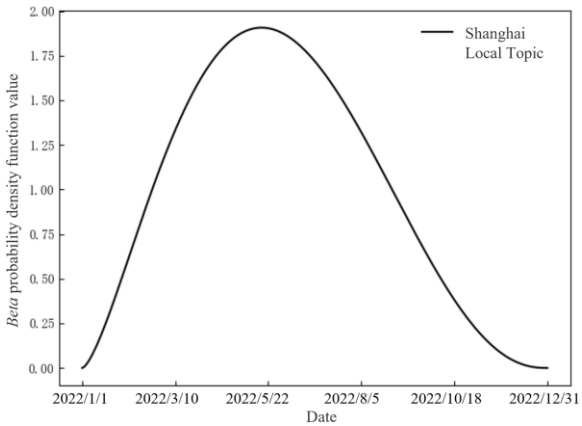

Figure 18. Time Varying Curve of Topic in Shanghai.

4. Conclusions

The fundamental innovation of ST-BTM is its design tailored to the spatio-temporal characteristics inherent in social media short texts. By jointly leveraging textual content along with corresponding temporal and geographic metadata, the model achieves higher topic coherence scores compared to conventional topic models under the same number of topics. Additionally, the topics discovered exhibit greater semantic consistency within each topic and effectively capture both the temporal evolution and spatial heterogeneity of topic distributions. Through rigorous experimentation, ST-BTM demonstrates superior coherence, meaningful topic differentiation, and robust performance across varied datasets.

This paper designs a Local Topics acquisition scheme. The document set in the same region is divided through the similarity between documents and Global Topics, and the appropriate document division threshold is determined with the help of the topic consistency index and the topic difference index. Then, accurately extract Local Topics from the complex document collection.

At present, the spatio-temporal topic model proposed in this study only considers the location and temporal information in social media check-in data. However, social media data also contains rich user-related information, such as user relationships. By incorporating more of this additional information, it is possible to uncover even more nuanced and diverse topic content. In this study, documents from different regions are classified into types based on similarity, successfully obtaining corresponding local topic results. However, this process is relatively complex. Future research could design a more straightforward and efficient Local-Global topic model that iteratively optimizes both global and local topic results, thereby simplifying the process of extracting local topics and improving accuracy.

References

Garcia, K., Berton, L., 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, 101, 107057.

Gangwar, P., Singh, V., 2021. Time-Based aggregation for Bi-term Topic Model to Analyze CoVID-19 Twitter Data. *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, Dehradun, India, pp. 1– 5.

Ghasiya, P., Okamura, K., 2021. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access*, 9, 36645–36656.

He, T., Yin, H., Chen, Z., Zhou, X., Sadiq, S., Luo, B., 2017. A Spatial-Temporal Topic Model for the Semantic Annotation of POIs in LBSNs. *ACM Trans. Intell. Syst. Technol.*, 8, 1–24.

Li, R., Lv, S., Zhu, H., Song, X., 2020. Spatial-Temporal Topic Model for Cold-Start Event Recommendation. *IEEE Access*, 8, 214050–214060.

Liu, H., Ge, Y., Zheng, Q., Lin, R., Li, H., 2018. Detecting global and local topics via mining twitter data. *Neurocomputing*, 273, 120–132.

Liu, Y., Ester, M., Hu, B., Cheung, D.W., 2015. Spatio-Temporal Topic Models for Check-in Data. *2015 IEEE International Conference on Data Mining (ICDM)*, IEEE, Atlantic City, NJ, USA, pp. 889– 894.

Mei, Q., Liu, C., Su, H., Zhai, C., 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. *The 15th International World Wide Web Conference 2006*, ACM, Edinburgh, Scotland, pp. 533– 542.

Qiang, S., Wang, Y., Jin, Y., 2017. A Local-Global LDA Model for Discovering Geographical Topics from Social Media. *Web and Big Data: First International Joint Conference, APWeb-WAIM 2017*, Springer International Publishing, Beijing, China, July 7– 9, 2017, pp. 27– 40.

Zhang, Y., Cai, X., Fry, C.V., Wu, M., Wagner, C.S., 2021. Topic evolution, disruption and resilience in early COVID-19 research. *Scientometrics*, 126, 4225–4253.