International Journal of Geo-Information

MDPI

Article

Evaluation of the Optimal Topic Classification for Social Media Data Combined with Text Semantics: A Case Study of Public Opinion Analysis Related to COVID-19 with Microblogs

Qin Liang [1], Chunchun Hu [1,*] and Si Chen [2]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430070, China; liangqin@whu.edu.cn
[2] Wuhan Natural Resources and Planning Information Center, Wuhan 430070, China; css@zrzyhgh.wuhan.gov.cn
* Correspondence: chchhu@sgg.whu.edu.cn

Abstract: Online public opinion reflects social conditions and public attitudes regarding special social events. Therefore, analyzing the temporal and spatial distributions of online public opinion topics can contribute to understanding issues of public concern, grasping and guiding the developing trend of public opinion. However, how to evaluate the validity of classification of online public opinion remains a challenging task in the topic mining field. By combining a Bidirectional Encoder Representations from Transformers (BERT) pre-training model with the Latent Dirichlet Allocation (LDA) topic model, we propose an evaluation method to determine the optimal classification number of topics from the perspective of semantic similarity. The effectiveness of the proposed method was verified based on the standard Chinese corpus THUCNews. Taking Coronavirus Disease 2019 (COVID-19)-related geotagged posts on Weibo in Wuhan city as an example, we used the proposed method to generate five categories of public opinion topics. Combining spatial and temporal information with the classification results, we analyze the spatial and temporal distribution patterns of the five optimal public opinion topics, which are found to be consistent with the epidemic development, demonstrating the feasibility of our method when applied to practical cases.

Keywords: LDA; topic model; BERT; topic classification; public opinion analysis

Citation: Liang, Q.; Hu, C.; Chen, S. Evaluation of the Optimal Topic Classification for Social Media Data Combined with Text Semantics: A Case Study of Public Opinion Analysis Related to COVID-19 with Microblogs. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 811. https://doi.org/10.3390/ijgi10120811

Academic Editor: Wolfgang Kainz

Received: 27 September 2021
Accepted: 28 November 2021
Published: 30 November 2021

1. Introduction

In the era of big data, the number of netizens has been increasing annually [1]. Social media provides important platforms for Internet users to express and exchange their views, as well as to obtain information. Once a significant event occurs, the public tends to describe their attention to and cognition of the event on social media platforms, leading to dissemination and discussion of the topic. In particular, from the end of 2019 to April 2020, the discovery of COVID-19 and lockdown policy prompted the public to exchange information on social media platforms in order to learn about the epidemic [2]. User-generated content (UGC), generated on some social media platforms such as WeChat, Sina Weibo, and Twitter, has become an important source for obtaining social sentiment and analyzing public opinion. Moreover, topic classification has a pivotal role in public opinion analysis, on which a considerable amount of literature [2–6] has been published regarding COVID-19. Although the value of revealing public opinions and emotions through social media has been proven in extensive research [7–9], the following challenges still exist. First of all, there are two sources of access to geotagged posts: one is from user registration information, which is coarse-grained, while the other is from check-in posts, which are fine-grained and shared by users initiatively, but only by a minority. Secondly, how to effectively mine valuable topic information from word-limited and casual-content social media posts requires further research. Finally, obtaining the appropriate classification of topics is one of the most frequently stated problems when using probabilistic topic models

for topic mining, as the number of topics affects the generation results of the model. In spite of many proposed researches [10], which were employed to assess the optimal number of topics but did not perform well in their consistence with people's subjective perceptions, the lack of semantic interpretability for topics has existed as a classification problem for many years. Therefore, it remains a challenge to obtain public opinion topics, including their spatial and temporal distribution, at a fine-grained scale within a city.

With the aim to analyze topics within online public opinion for a city during the COVID-19 epidemic, we propose a method that can address some of the problems mentioned above. First, we establish the correlation between the ground truth data related to the epidemic and check-in posts of social media, then reveal whether it is feasible to use only check-in posts for public opinion analysis. More than that, an algorithm for evaluating the optimal topic classification of public opinions is proposed, then the temporal and spatial distribution of various topics, based on the optimal classification results, is further discussed.

For this paper, a standard corpus was adopted to test the proposed evaluation method, and check-in posts on Weibo were applied to solve practical problems arising during the COVID-19 epidemic. Check-in posts are one of the geotagged post types in Weibo, which contain rich text, temporal, and spatial information. Based on check-in microblogs in Wuhan from December 2019 to April 2020, COVID-19 related posts were extracted to explore the relationship between check-in posts and the epidemic. Then, our evaluation method was proposed in order to find the optimal number of topics when the LDA topic model is used to classify public opinions, which substitutes the numerical expression for the subjective experience from a semantic point of view using a BERT pre-training model [11]. Finally, we analyzed the generated topics and characteristics of their temporal and spatial distribution according to the check-in data in order to provide a reference for the fine monitoring management and scientific governance of public opinion.

The remainder of this paper is structured as follows: Section 2 reviews the related studies in the literature. Section 3 explains the proposed method and other related methods. In Section 4, we compare our evaluation method with the other four using a standard corpus, then apply our approach to the COVID-19 case study and discuss the results. Section 5 provides our conclusions.

2. Related Works

Social media platforms, such as Weibo, Twitter, and Facebook, contain a variety of user-generated content (e.g., text, pictures, locations, and videos), making them content-rich and more convenient than traditional questionnaire surveys when obtaining a mass of research data. Research on social media has also become more and more popular. Some research works [12–15] have utilized social media data to analyze public opinions, with a focus on natural disasters and public health emergency. With the outbreak of COVID-19, social media has quickly become an important platform for information generation and dissemination. Topic mining, sentiment analysis, temporal and spatial distribution of topics, and similar perspectives for public opinion analysis have been mainly discussed among the many relevant studies in the literature [3,6,16,17]. Han et al. [6] combined the LDA topic model with a random forest model to classify COVID-19-related posts on Weibo, then analyzed the temporal and spatial distribution of different topics at the national scale. Their results showed that the variation trend of topics was in sync with the development of the epidemic over time. In addition, the spatial distribution of different topics was relevant to various factors, including disease severity, population density, and so on. Chen et al. [17] used SnowNLP and K-means methods to carry out sentiment analysis and topic classification, respectively, in the use of posts on Weibo under the epidemic situation, and verified the correlation between netizen emotions and the epidemic situation in various places. Sakun Boon-Itt and Yukolpat Skunkan [18] identified six categories of tweet topics by LDA topic modeling based on the highest topic coherence but dismissed the temporal

and spatial information. Han Zheng et al. [5] analyzed Twitter tweets by LDA to uncover temporal differences in nine topics that were identified by the existing methods [19,20].

From the perspective of data sources, existing researches [6,21] have adopted administrative districts, defined from user registration information, as the analysis unit while analyzing the spatial distribution pattern of topics or sentiments with social media data. Such information ignores the value of check-in posts on social media, with which we were able to reveal the distribution characteristics of public opinion topics within a city.

From the perspective of analytical methods for public opinion topics, the Latent Dirichlet Allocation (LDA) topic model [22] is considered to have the advantage of identifying topics from massive text collections in an unsupervised manner [23], and is also one of the most widely used probabilistic topic models. Wang et al. [24] adopted LDA to determine the latent emergency topics in microblog text and the corresponding topic–word distribution. On this basis, they used the SVM method to classify the new posts according to the appropriate topics, which provided decision support for the emergency response. Wang and Han [2,6] utilized LDA to initially set text classification labels, then introduced a random forest model to obtain the classification result for public opinion topics. Amina Amara et al. [25] exploited multilingual Facebook corpus to track COVID-19 trends with topics extracted by LDA topic modeling. However, those works did not provide a theoretical basis for the optimal number of topics in LDA. Numerous topic mining and recognition studies [24,26] have shown the practicality and effectiveness of the LDA model, but the number of topics in the model can affect the classification results [27].

By the means of introducing the concept of a topic, the LDA model can display text information in a topic space of lower dimension, leading to a good effect in text topic mining. With the aim to choose the appropriate number of topics, many methods have been proposed [19,20,28,29] besides the traditional methods, such as calculating the perplexity [22]. However, these methods only performed well for the model theoretically, and may not necessarily extract practical topic information consistent with experience. Under normal circumstances, the number of topics is obtained by experience or repeated experiments, which could lead to large errors [30]. In addition to experience-based selection approaches [31], perplexity-based approaches [22], Bayesian statistics methods [28], and the HDP method [32] are other classical methods used to obtain the appropriate number of topics in LDA; however, these methods are characterized by some problems, such as high time complexity or a lack of logical derivation.

Other methods [27] have been adopted to choose the optimal number of LDA topics, mainly based on the similarity between topics. Cao Juan et al. [19] utilized the cosine similarity to describe the stability of topic structure according to the topic–word probability distribution matrix generated by LDA, while Deveaud et al. [20] adopted the KL divergence to measure the similarity between probability distributions. Krasnov and Sen [10] proposed a clustering approach with a cDBI metric to assess the optimal number of topics, but this only worked well on a small collection of English documents. There have been some researches that tried to identify the optimal number of topics from a perspective of statistical physics. Ignatenko and Koltcov et al. [33] proposed the fractal approach to find out the optimal number of topics in a three topic modeling algorithm, i.e., PLSA, ARTM, and LDA models. With the assumption that the transition region corresponds to the "true" number of topics, they identified a range of figures, instead of a firm answer, for the optimal number of topics. Koltcov [34] regarded the number of topics as an equivalent of temperature in nonequilibrium complex systems (i.e., topic model). By calculating Rényi and Tsallis entropies based on a value of free energy of such systems, the optimal number of topics could be identified. These research works [19,20,33,34] chose the optimal number of topics based on the probability distribution of the LDA model but did not consider the semantic relevance of the topics generated by LDA, which may contribute to improving human interpretability of topics. LDA is based on the bag-of-words model, which ignores the contextual information between texts. Therefore, Wang Tingting et al. [27] improved the topic–word matrix generated by the LDA model and employed a topic–word vector matrix

by adding the semantic information of words using the Word2Vec word embedding model. Finally, the pseudo-F statistics in adaptive clustering were used to obtain the optimal number of topics. However, the Word2Vec model obtains a static word vector and ignores the contextual information of words. Different from the above research works, in this paper, we propose a new evaluation method, which considers the semantic similarity by combining a BERT pre-training model in order to obtain the optimal topic number for LDA.

3. Methods

3.1. LDA Topic Model

The Latent Dirichlet Allocation (LDA) topic model is an unsupervised machine learning algorithm proposed by Blei [22] in 2003, which adds Bayesian prior information to Probabilistic Latent Semantic Analysis (pLSA), and which can be used to identify potential topic information in a large-scale corpus.

LDA is a three-level hierarchical Bayesian model, which introduces the concept of a topic into a document–word matrix, then extends it to the combination of a document–topic probability matrix and a topic–word probability matrix. This model is based on the theory that similar words may belong to the same topic [35]. It assumes that a document is composed of multiple topics, and each topic is characterized by a distribution over words; that is, topics are latent. The framework of the LDA topic model is presented in Figure 1, where the box implies the number of repeated sampling steps, M represents the number of documents in the corpus, K represents the number of given topics, the number of words in the vocabulary list is denoted by N, and Z means the potential topics of the only observable words W. The LDA model chooses the topic for each document from the document–topic probability distribution θ (with size M × K; α is the hyperparameter of the Dirichlet prior distribution for the topic distribution underlying each document), then chooses the words under each topic from the topic–word probability distribution φ (with size K × N; β is the hyperparameter of the Dirichlet prior distribution for the word distribution under each topic), in order to generate the document.

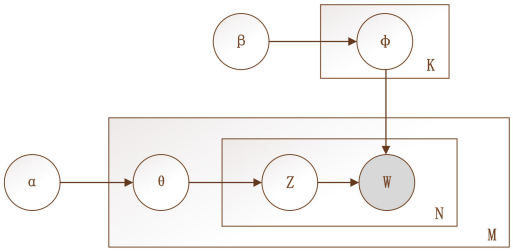


Figure 1. Schematic of LDA topic model.

In short, the LDA model consists of two processes:

① $p(\vec{z})$: The topic probability distribution $\vec{\theta}_m$ ($\vec{\theta}_m \sim Dir\left(\vec{\alpha}\right)$) of the mth document is obtained from the document–topic probability distribution, then the topic $Z_{m,n}$ of the nth word is generated.

② $p\left(\vec{w}\middle|\vec{z}\right)$: The word probability distribution $\vec{\varphi}_k$($\vec{\varphi}_k \sim Dir\left(\vec{\beta}\right)$) of topic k = $Z_{m,n}$ is obtained from the topic–word probability distribution, then the word $W_{m,n}$ is generated.

The goal of the LDA model is to identify the underlying topics characterized by words.

$$p\left(\vec{z}\middle|\vec{w}\right) = \frac{p\left(\vec{w},\vec{z}\right)}{\sum_z p\left(\vec{w},\vec{z}\right)} = \frac{p\left(\vec{w}\middle|\vec{z}\right)\cdot p(\vec{z})}{\sum_z p\left(\vec{w},\vec{z}\right)} \quad (1)$$

During the iterative process of the LDA model, only the topics of words are reallocated, while the document–topic and topic–word probability matrices are calculated using the

expectation formulas in the process of iteration [36]. The original document–word matrix, on which LDA is based, is built using the bag-of-words model. The bag-of-words model ignores the correlations between words and characterizes the document as a series of independent words directly, which leads to the problem of insufficient text semantics. At the same time, the selection of the number of topics affects the topic recognition ability of the LDA model.

3.2. Evaluation Methods to Determine the Optimal Number of Topics

A smaller number of topics leads to indistinguishable overlap between topics, while a larger number generates overly specific topics; thus, each topic may contain less information. The number of topics is a crucial parameter for the LDA model. In order to determine this parameter, scholars have proposed a variety of methods to evaluate the optimal number of topics. This paper does not repeat the basic methods, such as the perplexity approach, but mainly describes the following four evaluation methods that were commonly employed in recent works [27,37,38]. K denotes the number of topics in the rest of this section.

(1) Griffiths (2004) [28] adopted a Bayesian statistical method to determine the optimal number of topics, which means that the posterior probability of the model was used to obtain the optimal solution. In the LDA model, this method considers that the number of topics K corresponding to the maximum value of log P(w | K) is the best, where P(w | K) can be estimated by the harmonic average of P(w | z,K) under different topics.

(2) Cao Juan (2009) [19] utilized the nature of topic relevance to determine the most appropriate number of topics in LDA. In this paper, two definitions are given. The first is topic density, which means that the number of topics can be obtained when the cosine distance between these and other topics is less than a certain threshold. The other is model cardinality, which indicates the optimal number of topics when the topic density is less than a certain threshold. The basic idea of this method is as follows: First, initialize the value of K, then calculate the model cardinality C of LDA. Then, update the value of K constantly, according to the difference between C and K, until the average cosine distance and cardinality of the LDA model converge.

$$ave\_dis(structure) = \frac{\sum_{i=0}^{K}\sum_{j=i+1}^{K} corre(T_i, T_j)}{K \times (K-1)/2} \tag{2}$$

However, in this model, the vector representation of each topic is obtained according to the topic–word probability distribution matrix, and is only the probability distribution over words without considering semantic information.

(3) Arun (2010) [29] proposed a robust method to determine the correct number of topics considering both the document–topic matrix M2 and the topic–word matrix M1. The singular values, σ_i, derived from the singular value decomposition of the topic–word matrix represent the variance distribution of the topic. If the topics are well-separated, then σ_i should be equal to the L2 norm of the ith row vector of the topic–word matrix, $\forall i = 1, \ldots, K$. This method suggests a KL divergence measurement, as calculated by Equation (4). The larger the value of Equation (3), the better the LDA model performs.

$$ProposedMeasure(M1, M2) = KL(C_{M1}||C_{M2}) + KL(C_{M2}||C_{M1}) \tag{3}$$

$$KL(C_{M1}||C_{M2}) = \begin{array}{l} \sum_{i=1}^{K} C_{M1}(i) * \log(C_{M1}(i)/C_{M2}(i)) \\ + \sum_{i=1}^{K} C_{M2}(i) * \log(C_{M2}(i)/C_{M1}(i)) \end{array} \tag{4}$$

where C_{M1} is the distribution of singular values of the matrix M1, and C_{M2} is the distribution obtained by normalizing the vector L $*$ M2 (L is a $1 * D$ vector of lengths of each document in the corpus).

(4) Different from Arun and Cao, who selected the optimal number of topics according to the similarity or distance between topics, Deveaud (2014) [20] proposed a simple heuristic

method by maximizing the information divergence, D, between topics. T_k in Equation (5) is the set of K topics modeled by LDA. The information divergence D is calculated using the JS divergence, according to Equation (6), in which W_k is the set of n words that have the highest probabilities $P(\omega|k)$ in topic k.

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{(k,k')\in T_K} D(k||k') \tag{5}$$

$$D(k||k') = \frac{1}{2} \sum_{\omega \in W_k \cap W_{k'}} P(\omega|k) log \frac{P(\omega|k)}{P(\omega|k')} + \frac{1}{2} \sum_{\omega \in W_k \cap W_{k'}} P(\omega|k') log \frac{P(\omega|k')}{P(\omega|k)} \tag{6}$$

3.3. Evaluation Method for the Optimal Topics Combined with Text Semantics

In practical applications, most of the topics obtained based on the statistical model are the optimal results for the model, while the underlying topics may not have practical significance. Therefore, experience-based evaluation methods for determining the number of topics are still applied in various studies [27]. Under this circumstance, we choose the optimal number of topics according to the text semantic information of the topic, which is characterized as the distribution over words generated by LDA. Essentially, we construct an evaluation index to express the experience-based method objectively.

In an optimal topic structure, each topic should be a compact, meaningful, and interpretable semantic cluster, and topics should be mutually exclusive [19]. With reference to the clustering evaluation index, this article is based on the idea that "the higher the word semantic similarity within topics is, the lower the word semantic similarity between topics is". According to the topic–word (T–W) probability matrix generated by LDA, we consider each topic as a sentence made up of words in descending order of probability, which can be obtained from the distribution. Then, we calculate the text semantic similarity to choose the optimal number of topics. Text similarity is measured by the classical cosine similarity method. In terms of text vectorization, we considered the excellent characteristics of the BERT pre-training model in natural language processing [11], which can fully take into account the context relations of words in documents and the lexical polysemy.

Therefore, we adopted a BERT model to realize text vectorization. In order to obtain the optimal classification results, the RI index was constructed for evaluation of the quality of the classification results. Assuming that the number of topics generated by LDA is K, and each type of topic consists of N words, then the RI value is calculated as follows

$$RI(K) = \frac{\sum_{k=1}^{K} CS_{intra_k}/K}{CS_{inter}} \tag{7}$$

$$CS_{intrak} = \frac{2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \cos\_sim\left(W_{ki}, W_{kj}\right)}{N(N-1)} \tag{8}$$

$$CS_{inter} = \frac{2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} \cos\_sim\left(S_i, S_j\right)}{K(K-1)} \tag{9}$$

where the *cos_sim()* function in Equations (8) and (9) represents the cosine similarity between two vectors, which is normalized between 0 and 1; W_{ki} represents the word vector of the ith word under topic k in the T–W matrix; S_i represents the sentence vector composed of the words under the ith topic. As a topic generated by LDA is composed of words, the number of words—denoted by N—will affect the classification result. It is necessary to calculate the value of RI under different values of N, finally taking the mean value as the index value to choose the optimal number of topics. A larger value of RI implies a better topic structure.

$$\hat{K} = \underset{K}{\operatorname{argmax}} RI(K) \tag{10}$$

The whole algorithm process is depicted in Algorithm 1, which can be divided into the following steps:

(1) Given the number of topics (denoted by K), execute the LDA model, and obtain the document–topic and topic–word probability distribution matrices.
(2) Set the number of words N, which is meaningful when characterizing the topic as a series of words. Then, each topic is represented by the first N words with the largest probability values in the topic–word matrix.
(3) The words of the ith (i = 1, 2, . . . , K) topic are connected into sentences, and the BERT model is used to obtain the sentence vector of the ith topic.
(4) The word vector of each word under the ith topic is obtained using the BERT model, and the cosine formula is used to calculate the text similarity between words. Then, the mean value is calculated as the intra-cluster similarity of the ith topic.
(5) Repeat steps (3)–(4) until i = K. Then, calculate the similarity between the words of two topics and take the mean value as the inter-cluster similarity, and take the mean value of the intra-cluster similarities as the intra-cluster similarity for all topics. Finally, the RI value is obtained by taking the ratio of intra-cluster similarity to inter-cluster similarity.
(6) Choose different values of N and repeat steps (2)–(5).
(7) Choose different values of K and repeat steps (2)–(6).

Algorithm 1: How to get RI values under different number of topics K

Data: Document corpus after preprocessing, the number of topics K
Result: RI values under different number of topics K
for $K \leftarrow 2$ **to** 20 **do**
  run **LDA** topic model to get **D-T** and **T-W** matrices;
  Input: T-W matrix of size $K \times N$, the number of words N that represent each topic
  for $N \leftarrow 5, 10, 15, 20$ **do**
    for $i \leftarrow 1$ **to** K **do**
      the ith row of T-W matrix makes up a sentence i;
      use BERT to obtain sentence embedding S_i
      for each *word of sentence i* **do**
        use BERT to obtain word vectors matrix WV with the size of $N \times 768$;
        compute the pairwise cosine similarities between all rows in WV;
        compute the mean value of cosine similarities CS_{intra}
      end
    end
    compute the cosine similarities between each sentence vector S_i
    then get the average value CS_{inter}
    compute RI value under K, i.e. RI=CS_{intra}/CS_{inter}
  end
end

4. Experiment and Results

Wuhan city, which belongs to Hubei province, is one of the regions most severely affected during the early days of the COVID-19 outbreak. Analyzing public opinion within the city is helpful to understand the status of the epidemic in this region and the public opinion trend, which has important guiding significance for differentiated epidemic prevention and control in the future. We first evaluated the validity of the proposed method, which can be used to choose the optimal number of topics. After that, we carried out the topic mining experiment based on the real case study, analyzing the spatio-temporal distribution characteristics of different topics under the optimal classification results. The overall workflow of our experiments is shown in Figure 2. The CPU used in our experiment was an 11th Gen Intel(R) Core (TM) i5-1135G7 CPU @ 2.40 GHz and the memory was

16.0 GB of RAM. The LDA model was conducted with the scikit-learn Python module [39]. It should be noted that the largest part of our experiment was realized in Python language, except for the comparison results of four methods mentioned in Section 3.2, which were performed in R language.

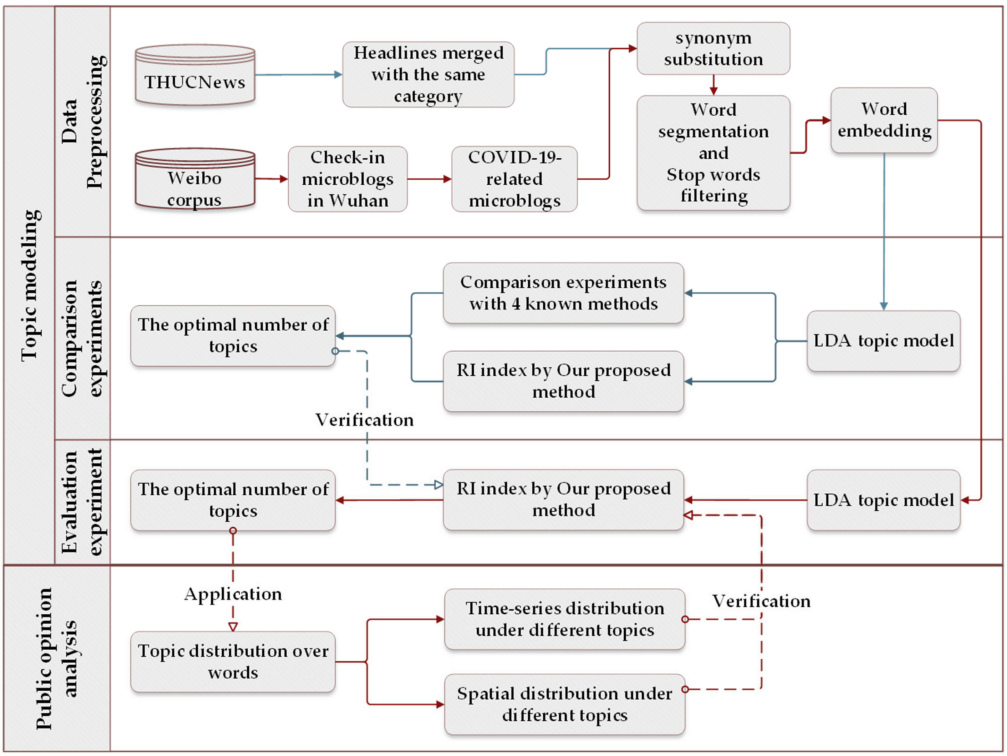


Figure 2. The workflow of experimental implementation.

4.1. Comparison Experiments Based on the Standard Corpus

In order to verify the effectiveness of our method, we employed a standard classified Chinese news corpus to conduct experiments. As the content of posts on Weibo is limited to 140 characters, we extracted nine categories of news headlines from the Tsinghua University news corpus (THUCNews) [40] and merged the news headlines under the same topic category randomly, in order to form 162,000 documents having 140 characters or less. Some instances of this dataset are presented in Table 1. With reference to [21], synonym substitution was carried out during the preprocessing.

With regard to the four evaluation methods described in Section 3.2, the R language package ldatuning [41] was used to implement the experiment of finding the optimal number of topics for this corpus. The experimental results are shown in Figure 3, where the dotted line indicates the criteria for which a smaller index value implies better performance of the LDA. In contrast, the solid lines indicate the criteria for which a larger index value implies better performance of the LDA. The results demonstrate that the optimal number of topics obtained by different methods was not the same, and none of them found the standard number of news categories (i.e., nine).

Table 1. Some instances of the THUCNews dataset before the process of merging.

Category	Headline Text	Amount of Instances in the Category
finance	Fund online hit a situation exposure, the proportion and retail investors have no difference.	18,000
realty	The pain of high land price in Beijing property market: land costs account for more than 30% of the housing price.	18,000
education	Microblogs are a popular way for college entrance exam candidates to relieve stress.	18,000
science	Buyers will see: recommendations for mobile phones that worth buying in November	18,000
society	A woman has been looking for her husband for 14 years before she finds out he has a new family.	18,000
politics	US President Barack Obama has postponed the move of the US embassy in Israel.	18,000
sports	Maldini: AC Milan are far behind Real Madrid and Barcelona, Ibrobi is unlikely to start a new dynasty.	18,000
game	Theme Day Activity of online game "King of Kings 3" in Tencent version.	18,000
entertainment	Jay Chou responded to the poor ratings by saying that it is normal to lose sometimes	18,000

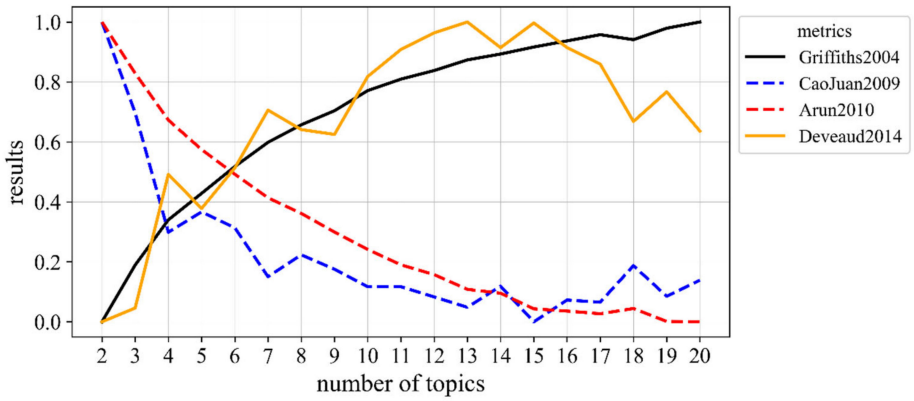


Figure 3. The comparison results for the optimal number of topics using four different metrics.

As the LDA model obtains latent topics represented by distributions over words, when the proposed method is used for experiments, the choice of the number of words under each topic can also have a certain impact on the RI value results. The RI index was calculated, as shown in Figure 4, for cases where the topic text was composed of different numbers of words.

Therefore, within a certain range (N = 5, 10, 15, or 20 were selected for testing in this paper), no matter how many words were chosen to represent topics, the optimal number of topics obtained by our method was 9, consistent with the actual situation of the corpus. At the same time, according to the document–topic matrix, 161,856 documents were classified under the correct category (Table 2), and the corresponding accuracy rate was 99.9%. These results demonstrate the validity of the proposed method when using a standard corpus and the feasibility of the LDA topic model under the premise of choosing the optimal number of topics.

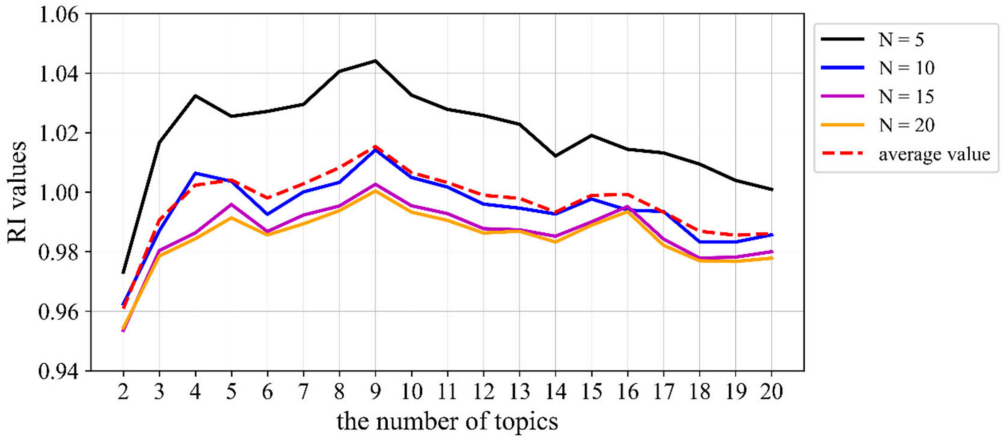


Figure 4. RI index results for our method based on the standard news corpus.

Table 2. LDA results for standard Chinese corpus with nine topic categories.

Category	Words That Describe the Topic	Actual Number of Documents	The Number of Documents Classified Correctly
finance	fund, future, goods, market, company	18,000	17,999
realty	price, opening, fine decoration, villa, Beijing	18,000	17,986
education	college entrance examination, graduate school exam, enrollment, examinee, offer	18,000	17,996
science	cellphone, Sony, Canon, internet, Nikon	18,000	17,937
society	man, woman, driver, ten thousand yuan, dead	18,000	17,989
politics	dead, president, happened, Obama, Iran	18,000	17,987
sports	Rocket, Barcelona, Real Madrid, Milan, player	18,000	17,996
game	game, online, online game, publish, player	18,000	17,981
entertainment	deny, expose, pose, star, respond	18,000	17,985

4.2. Case Study of Public Opinion Analysis during COVID-19 Epidemic in Wuhan

4.2.1. Data Sets and Preprocessing

The study area used in this paper includes 13 municipal districts of Wuhan, including Wuchang District, Hongshan District, and so on. The basic approach to obtain check-in data in Weibo is to use web crawler technology to capture the posts on the check-in webpage. Due to a permission limitation, users can only query the real-time check-in posts of about 20 pages of a certain place, which means that we could not crawl the historical records from this webpage, and the number of posts that we were able to crawl was limited. According to the user pool with 20 million active domestic and overseas users established by Yong Hu et al. [42], the authors crawled all posts of every user on Weibo and built a rich Weibo corpus. Based on the Weibo corpus provided by Yong Hu et al. [43], we filtered the check-in data with a location in Wuhan before and after the COVID-19 outbreak period in Wuhan; that is, from December 2019 to April 2020. The available attributes of data included: posting time, location (marked as latitude and longitude), and text. The daily number of confirmed cases in Wuhan was collected from the website of DingXiangYuan [44], among which the data of confirmed cases in December 2019 came from laboratory-confirmed data provided

in the literature [45]. The maps of Wuhan city, with basic elements and administrative areas, were obtained from the map world website. Considering the significant landmarks during the COVID-19 epidemic, we added three railway stations, Wuhan airport, and two special hospitals (Huoshenshan and Leishenshan) as data sources (Figure 5).

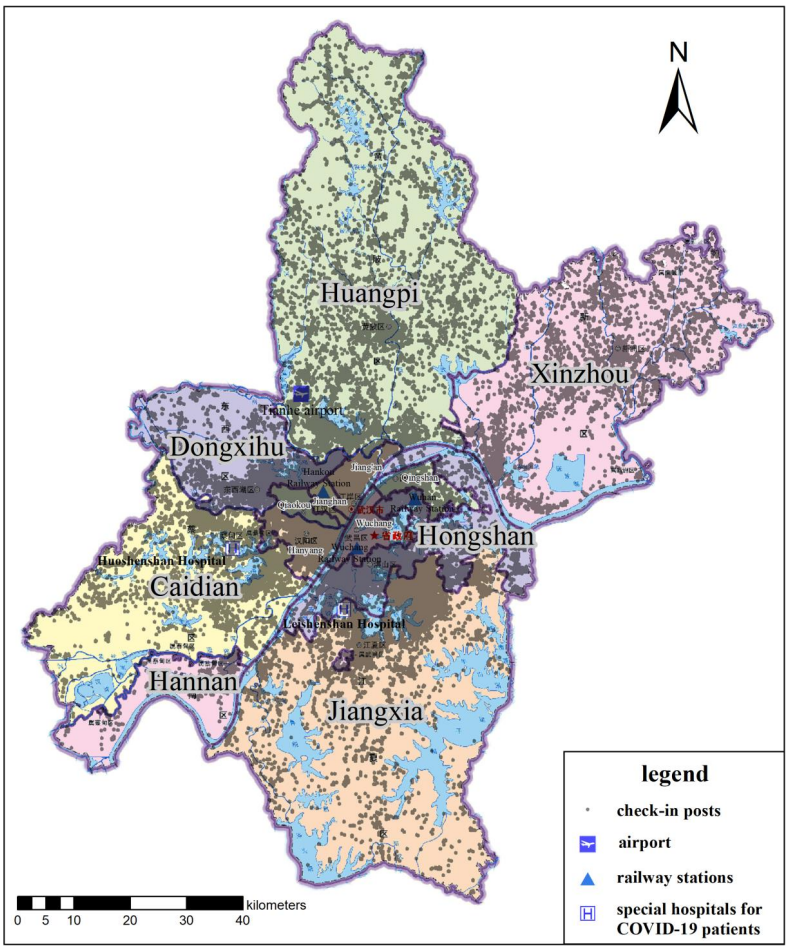


Figure 5. Administrative map of Wuhan with check-in posts on Weibo and other important POIs.

We first transformed the coordinates of the obtained check-in posts on Weibo in order to unify the coordinate system. Then, we extracted the check-in data in Wuhan, according to the spatial scope, and 617,032 check-in posts produced by 124,281 users were obtained. Some instances of this dataset were presented in Table 3. The posts distribution among the users is demonstrated by Figure 6a, from which we could discover several non-human contributors combined with text content. Sina Weibo users can share text messages limited to 140 characters, as well as non-text messages such as pictures and videos. The extraction process of text messages, pictures, videos, and other information will produce redundant text. In addition, hashtags and main text content will be inconsistent due to advertising and marketing. Therefore, this kind of text needed to be filtered. In general, text filtering includes the removal of interfering information, such as the names of the check-in places, links, @ symbols, pictures, videos, and hashtags. The initial filtered check-in text was screened, according to 179 epidemic-related keywords, such as "COVID-19", "pneumonia", "Cov-19", and "COVID-19" [43], after which 46,837 COVID-19-related posts were obtained. In order to carry out topic mining and reduce the interference of synonyms, word segmentation and stop word filtering were carried out and, after that, we filtered out the words with high frequency, such as "Wuhan", "epidemic", and "fighting". With reference to [21], synonym substitution was carried out (e.g., "Wuhan city" was replaced with "Wuhan", which both mean "Wuhan") and, finally, 46,774 non-empty posts produced

by 22,733 users were obtained. The posts' distribution among the users is demonstrated by Figure 6b, from which we could figure out that there were no non-human contributors combined with the text content. The characteristics of the datasets were shown in Table 4.

Table 3. Some instances of Weibo corpus obtained from the literature [43].

User_id	Created_at (GMT + 8)	Text Content	Check-in Location in GCJ-02 Coordinate System
cc44af5e7e03be03	31 December 2019 21:43:58	This morning's epidemic news didn't dampen Wuhan people's enthusiasm for New Year's Eve in Jiangtan. [ha-ha] Wuhan·Jiangtan; show map.	"114.298421,30.57753"
806ac40de73607d7	28 January 2020 15:40	The sun is shining for the first time since the lockdown of the city. I just want to clean my house. I really hope the epidemic will pass soon. Wuhan will win! #Go Wuhan# Wuhan; show map.	"114.200958,30.600012"
86a81fd176af326f	26 April 2020 19:01	For the better resumption of work and production in Wuhan, we will not rest [doge]. Wuhan· Wuhan Sixth Hospital; show map.	"114.28953,30.60014"

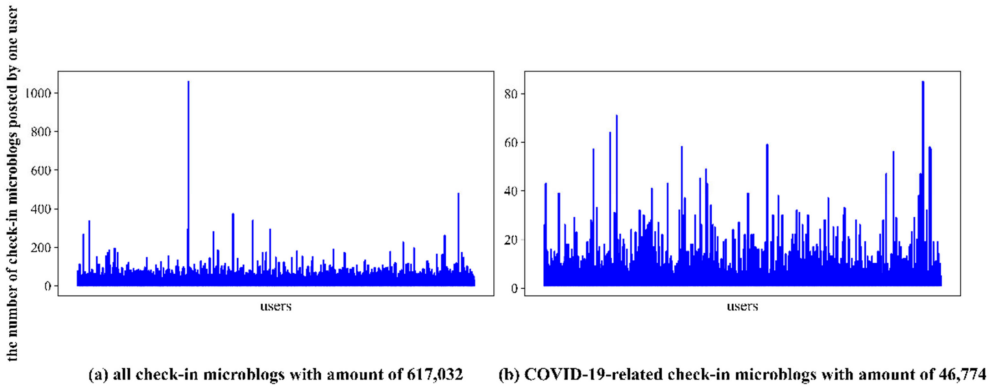


Figure 6. Distributions of the number of check-in microblogs among the users before and after data preprocessing.

Table 4. The characteristics of the datasets.

Datasets	Time Period (GMT + 8)	Space Range	Amount	Amount of Contributed Users	Fields
THUCNews	None	None	162,000	None	Category/headline text
All check-in microblogs	From 1 December 2019 0:00 to 30 April 2020 23:59	Wuhan city	617,032	124,281	User_id/time/location/text content
COVID-19-related check-in microblogs	From 1 December 2019 0:00 to 30 April 2020 23:59	Wuhan city	46,774	22,733	User_id/time/location/text content

4.2.2. Time-Series Analysis

Sina Weibo is one of the important ways for officials and the public to make their voices heard. The initial public opinion on COVID-19 originated from a widely spread official document, with the title "unexplained pneumonia," on Weibo. The more attention a topic gets, the more posts will be posted about it. In order to explore the relationship between public opinion on Weibo and the development of the epidemic, we first explored the correlation between the number of check-in posts and confirmed cases over time. Due to the order of magnitude difference between the number of confirmed cases and posts,

we adopted a logarithmic transformation method to reduce the absolute value of the obtained data—that is, n was converted to $\log_{10}(n + 1)$ after logarithmic conversion—with the purpose of shrinking the scale of data without changing the correlations among the data. The temporal process of the number of check-in posts and confirmed cases over time is shown in Figure 7, where the grey area represents the period (Φ) during which Wuhan was on lockdown (i.e., from 23 January 2020 to 8 April 2020; when the outbound travel restrictions from Wuhan were lifted).

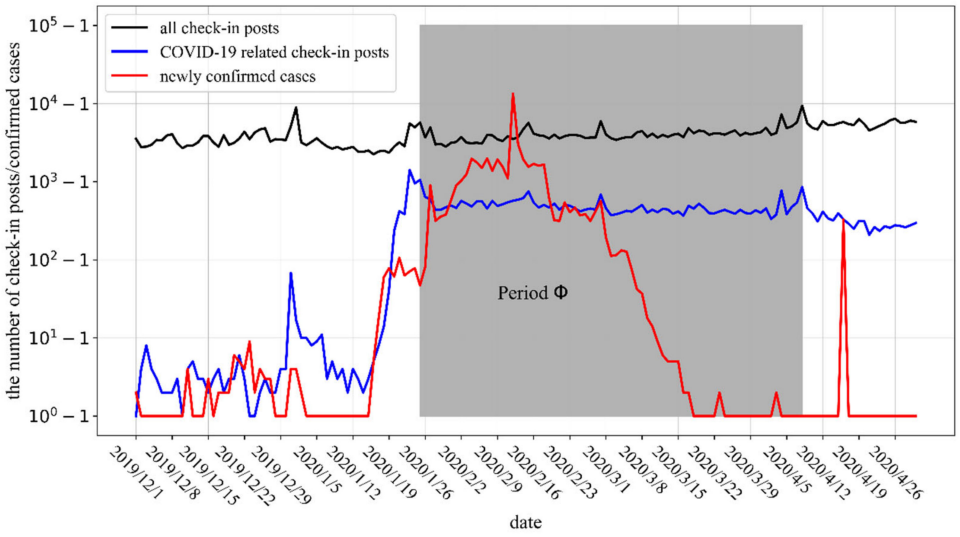


Figure 7. Time-series of check-in posts and new confirmed cases in Wuhan city.

Due to the fact that the number of active users on Weibo has had relatively little fluctuation in recent years [46], we found that the number of all check-in posts in Wuhan was relatively stable over time. During the New Year's holiday (from 30 December 2019 to 1 January 2020), the Spring Festival (from 22 to 28 January 2020), and Valentine's Day (on 14 February), there were small increases in the number of check-in posts, as well as when the rising trend of the epidemic in Wuhan was suppressed on 28 February, when the national mourning was held on 4 April, and when the outbound travel restrictions on Wuhan were lifted on 8 April, according to Figure 7.

The number of COVID-19-related check-in posts increased significantly on 31 December 2019. Then, they rapidly increased from 16 to 23 January 2020, and peaked on 23 January. In terms of the review of epidemic development, an unexplained pneumonia broke out in Wuhan at the end of 2019, the number of confirmed cases began to rise on 16 January 2020, and the city was put into lockdown due to coronavirus on 23 January 2020. Therefore, what can be clearly seen is that, during the outbreak of COVID-19, the quantitative change in COVID-19-related check-in posts roughly coincided with the outbreak timeline. As Wuhan city was on lockdown, the number of epidemic-related check-in microblogs (i.e., posts on Weibo) gradually decreased to a relatively stable number, and the number of all check-in microblogs maintained the same trend. Furthermore, the discussion about the epidemic occupied a large part of all of the check-in microblogs. As for the rapid growth in the number of newly confirmed cases on 16 April 2020, the direct reason is that the number of confirmed cases was revised in Wuhan city and, so, the number of check-in posts did not fluctuate significantly at the same time.

Spearman correlation analysis was conducted on the number of COVID-19-related check-in microblogs and the number of newly confirmed cases. The results showed that the global correlation coefficient was 0.56 with 99% confidence. Therefore, the variation trend of the number of check-in posts on Sina Weibo reflected the development of the epidemic situation, to a certain extent, and public opinion analysis based on check-in microblogs has a certain reliability.

4.2.3. Spatial Distribution of the COVID-19 Related Microblogs

To some extent, the variation trend of the number of check-in microblogs over time can reflect the development of the COVID-19 epidemic. Thus, check-in posts on Weibo can be used as a data source for public opinion analysis. In order to explore the spatial distribution characteristics of public opinion on Weibo further, the map of Wuhan city including basic elements was regarded as the base map, and we used the kernel density estimation method to explore the spatial distribution of COVID-19-related check-in microblogs.

Figure 8 shows the kernel density estimation results based on the check-in microblogs related to the epidemic. As the visual results show, the hot-spots of check-in microblogs were mainly distributed in the center of the main urban area of Wuhan city. In addition, there were other hot-spots scattered in the municipal districts around Wuhan (Huangpi District, Xinzhou District, Dongxihu District, Caidian District, Hannan District, and so on), as well as Tianhe Airport. The hot-spots of check-in microblogs in municipal districts were mainly distributed in residential areas. This was due to more densely populated places having more users on Weibo and, thus, more posts with fine-grained geo-tags (i.e., check-in microblogs) were uploaded. As the traffic main artery access to Wuhan, Tianhe airport and the three major railway stations had a relatively dense floating population during the period of non-lockdown, thus, also making them check-in hot-spots on Weibo.

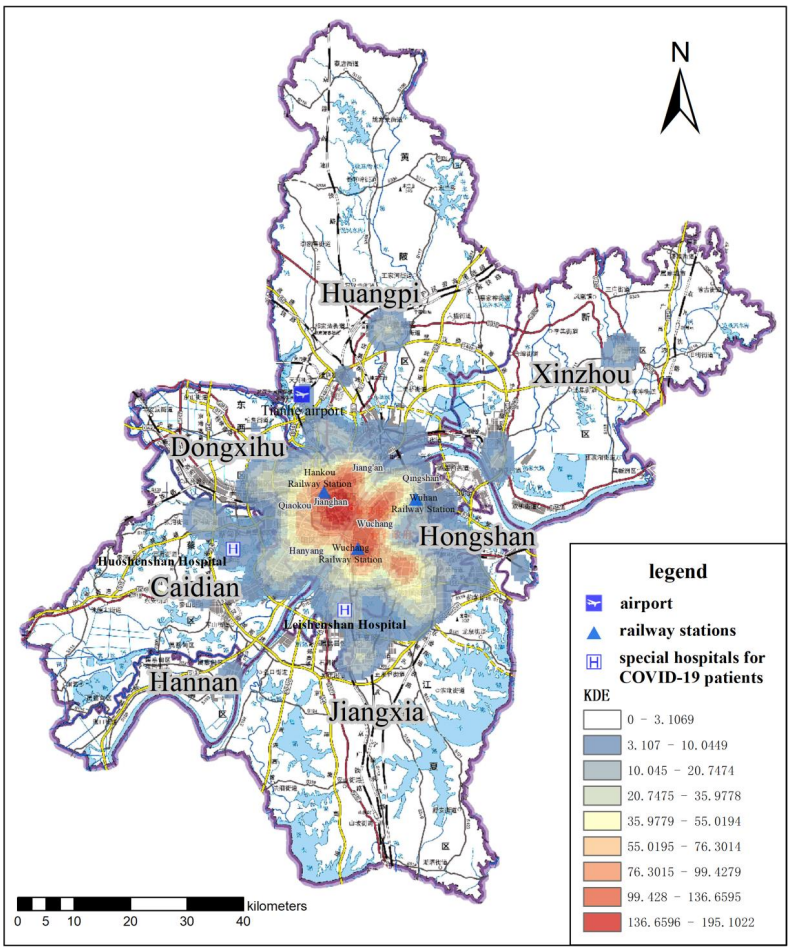


Figure 8. Kernel Density Estimation (KDE) results for COVID-19-related microblogs. (The KDE result was calculated with the bandwidth of 4405 m and the cell size of 529 m × 529 m.).

4.2.4. Evaluation Experiment on Determining the Optimal Number of Topics

The LDA topic model was adopted to mine the topics of the text documents on Weibo, by which the document probability distribution over topics and the topic probability distribution over words were obtained. According to the evaluation method devised in

this paper, we determined the optimal number of topics in LDA by using the bert-as-service Python module, developed by Xiao Han [47], to convert the words or sentences into fixed-length vectors (i.e., word/sentence embeddings). Then, the vector distance was characterized using the cosine similarity metric. We selected different amounts of words, N, to describe the topics, and obtained the results shown in Figure 9.

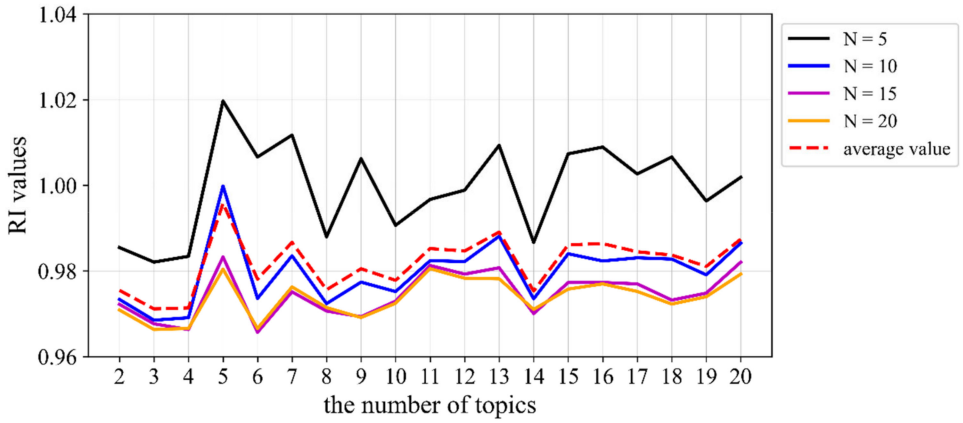


Figure 9. RI value results of our method based on the COVID-19-related corpus of Weibo.

As a result, we divided the epidemic-related check-in microblog corpus into five categories. The words that were used to describe each topic and the number of microblogs in each category are provided in Table 5. "Family care" refers to the public's concern and sympathy for their family and friends during the epidemic, and depicts the state of getting along with family members. "Home life" includes people's comments on community life and attention to commuting and traffic conditions. "Epidemic report" refers to reports of newly confirmed cases and other statistical data during the COVID-19 pandemic on Weibo, or objective comments on the status of confirmed cases. "Response status" refers to public concern about the current situation of the COVID-19-related response and treatment, including hospital capacity for nucleic acid testing, patient infection status, and community isolation. "Appreciation and praying" refers to the public's appreciation for frontline medical workers, blessings to the people of Wuhan, praise for those who prevent and control the epidemic, and hope for an early victory in epidemic prevention and control.

Table 5. LDA results of COVID-19-related check-in microblogs corpus with five topic categories.

ID	Topic Summary	Words That Describe the Topic	Number of Microblogs
1	Family care	Mom, at home, friend, family, dad, go home, child, life, work, on duty	11,912
2	Home life	Mask, work resumption, go out, community, lockdown, supermarket, lift the lockdown, at home, on duty, express	11,444
3	Epidemic report	Virus, confirm, case, new, pneumonia, COVID-19, China, infect, country, coronavirus	6918
4	Response status	Hospital, community, patient, quarantine, community, detect, nucleic acid, confirm, doctor, infect	6618
5	Appreciation and praying	Frontline, appreciate, city, people, medical workers, anti-epidemic, China, hero, early, national	9882

According to the document–topic probability distribution matrix, the topic category of each microblog was obtained, and the spatial and temporal statistical analysis (Figures 10 and 11) of microblogs under different topics was conducted.

The spatial patterns of public opinion topics 1–5 were aggregated and distributed, most of which were clustered within the Fifth Ring Road of Wuhan, while the hot-spots outside the Fifth Ring Road were mainly located in the residential areas in the centers of municipal districts and at Tianhe Airport. Therefore, we focused on the analysis of public opinion distribution within the Fifth Ring Road of Wuhan city. Figure 11 shows the spatial distribution maps and kernel density estimation results within the Fifth Ring Road of Wuhan City and the area near Tianhe Airport.

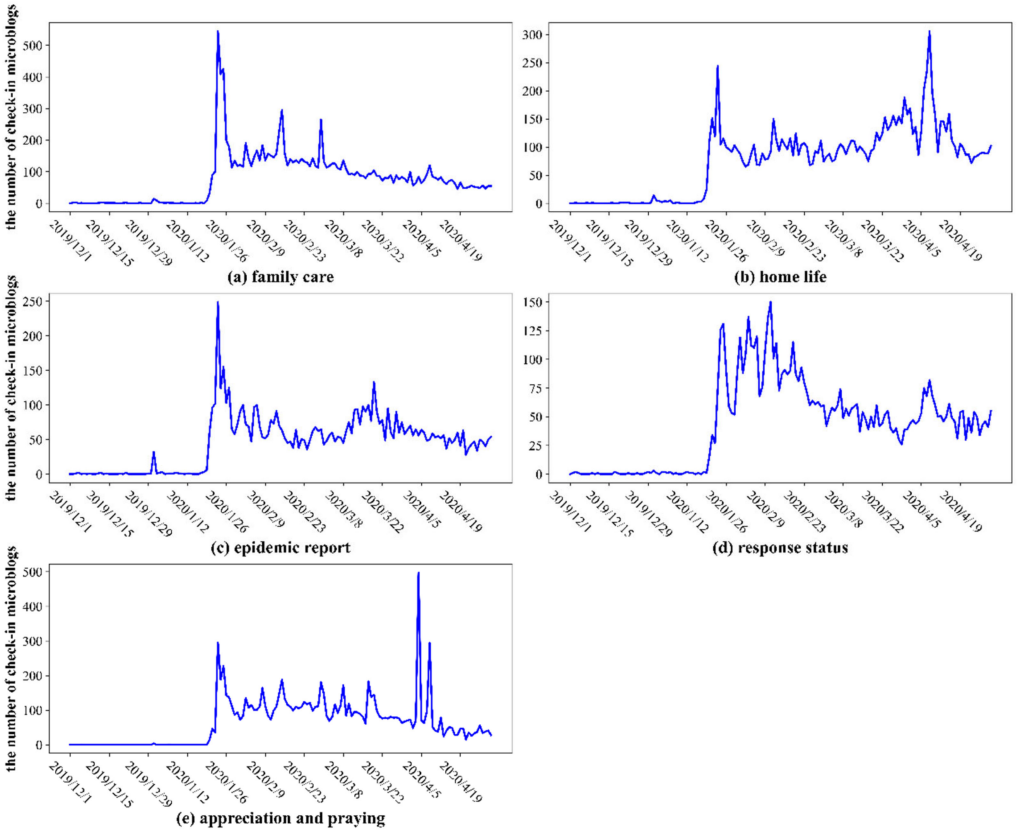
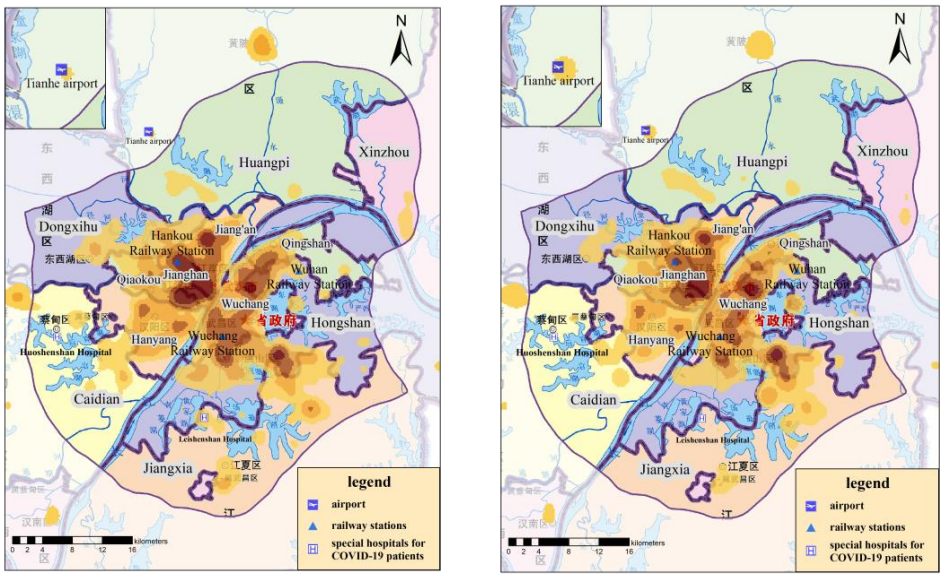
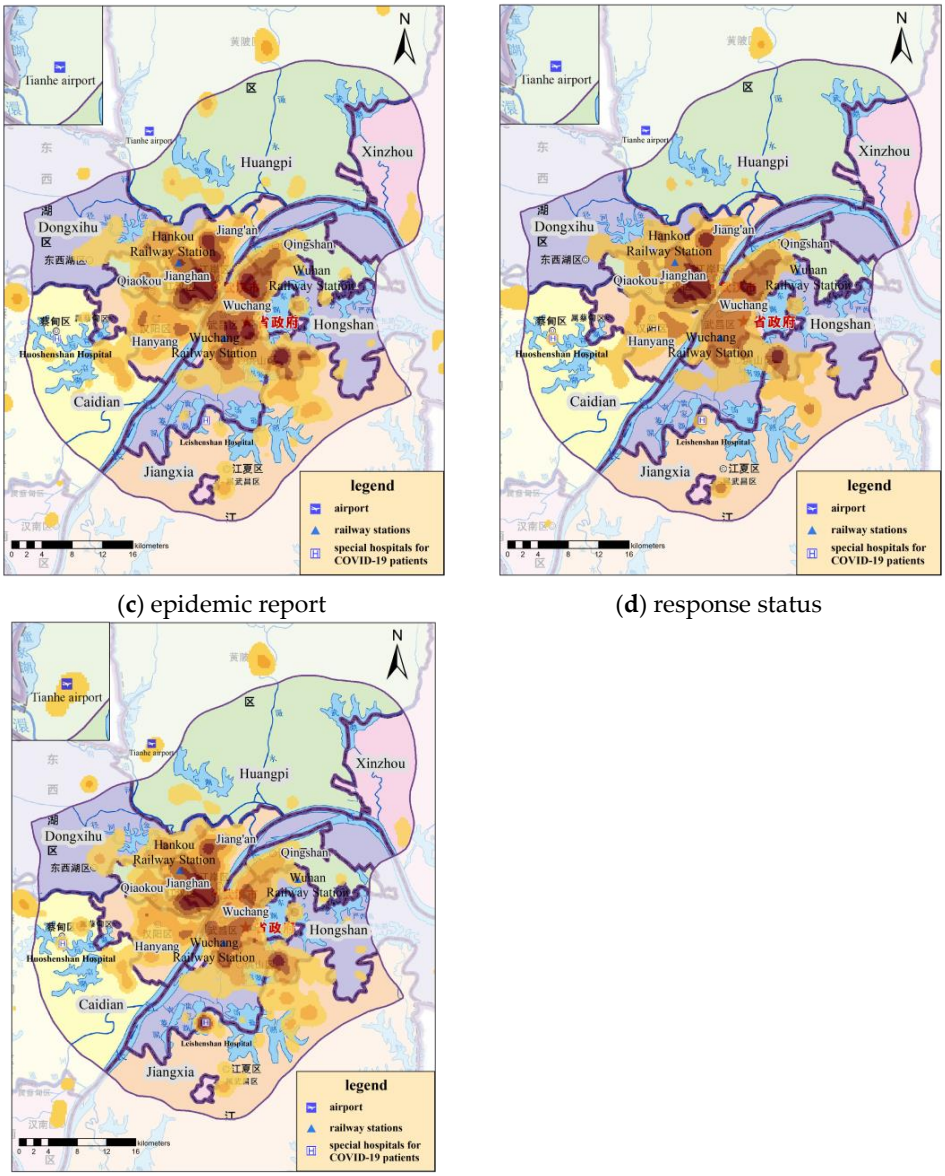


Figure 10. Time-series of the daily amount of check-in microblogs under different topic categories.



(**a**) family care (**b**) home life

Figure 11. *Cont.*



(**c**) epidemic report (**d**) response status

(**e**) appreciation and praying

Figure 11. KDE results of COVID-19-related check-in microblogs under different topic categories. (All KDE results were calculated with the bandwidth of 1000 m and the cell size of 200 m × 200 m).

4.3. Analysis and Discussion

4.3.1. Time-Series Analysis of Public Opinion Topics

The time variation trends of topics 1–5 are shown in Figure 10. The topic of "family care" reached its peak on the day of the lockdown on Wuhan (23 January), and remained at a high level in the following four days, due to the New Year's Eve and Spring Festival holidays, maintaining a relatively stable state since then. In addition, there was a slight increase around important time points, such as 2 February, on which the daily newly confirmed cases in Wuhan exceeded one thousand, Valentine's Day on 14 February, intercalary Day on 29 February, and 8 April, on which the lockdown of Wuhan ended. The topic of "home life" surged to a local maximum on the day of the lockdown, returned to a more stable state, then surged to a peak on the day of the lockdown ending (8 April). The "epidemic report" topic reached its peak on the day of the lockdown, then fell to a plateau and reached its relative maximum on 19 March, as all the newly confirmed and suspected cases were cleared the day before this day. As for the "response status" topic, several peaks occurred

in the early stage of epidemic prevention and control, and continued to fluctuate within the high value range. After that, it decreased with small fluctuations, and another local maximum appeared around the date that the lockdown lifted. The topic of "appreciation and praying" reached a high value on the day when the lockdown was implemented, then dropped to a low value and remained in a stable state of fluctuation. It reached a peak on 4 April—the National Day of Mourning—and also showed a peak on 8 April—the day when the lockdown was lifted.

Overall, all public opinion topics reached a peak on the day of lockdown; "Family care" increased around important time points; "Home life" was particularly focused on travel status, so there was a peak on the days of lockdown and lifting lockdown; "Epidemic report" reached another peak at the time of "double zero clearing"; to some extent, "response status" also reflected the public concerns about the epidemic, which fluctuated from the high value range in the early stage to the low values later, finally reaching another peak when the lockdown was lifted; "appreciation and praying" fluctuated steadily until it peaked on the day of national mourning and when the lockdown was lifted. Therefore, the variation trends of time-series under different topics were closely related to the category of text content, thus verifying the validity of the classification results.

4.3.2. Spatial Distribution of Topics of Public Opinion

Overall, the spatial distribution of each topic had similar characteristics, and the hot-spots of check-in microblogs for any topic were all in the downtown residential area of Wuhan. The topics of "family care", "home life", and "appreciation and praying" attracted different levels of public opinion near Tianhe Airport. "Family care", "home life", "response status", and "appreciation and praying" were mainly concentrated in areas with high population, such as in Jianghan District, while "epidemic report" was widely distributed through Wuchang District, Jianghan District, Hongshan District, and Jiang'an District. Analyzing the text content, farewell and nostalgia under the topic of "family care" appeared in traffic arteries, to a certain extent. Focusing on the willingness to travel within the "home life" topic and the description of the scene of seeing off medical staff in the airport within the "appreciation and praying" topic led to the emergence of public opinion hot-spots near Tianhe Airport. At the same time, hot-spots appeared near Leishenshan and Huoshenshan Hospitals under the topic of "appreciation and praying". Analyzing in combination with the time-series results, the topics of "home life" and "appreciation and praying" reached their peak the day before and after lifting lockdown, such that the related hot-spots were more scattered, compared to the other topics.

The popularity of "epidemic report" and "response status" topics near Tianhe Airport was relatively low. Considering the time-series results, this was partly due to the heated discussion of these two topics, which mainly occurred over the lockdown period, as the airport was located in the outskirts without a large residential area nearby. The other topic, "epidemic report", is mainly a statement of objective facts, so it was relatively evenly distributed in residential areas on both sides of the Yangtze River. "Response status" is closely related to medical resources, so hot-spots of public opinion appeared near both Huoshenshan Hospital and Leishenshan Hospital.

5. Conclusions

The analysis of online public opinion under the epidemic situation is of great significance for guiding public opinion and gaining access to public sentiment and public events. Utilizing a BERT pre-training model for word embedding, we proposed a semantic similarity-related evaluation method to finding the optimal number of topics generated by the LDA model, then adopted this method to analyze the online public opinion of check-in microblogs. We further analyzed the differences in the temporal and spatial distributions of the classified topics. We can draw the following conclusions:

(1) The text semantic-based evaluation method of finding the optimal number of topics is an objective representation of subjective experience, essentially, which can generate

the best number of explainable topics, to some extent, by considering semantic similarity between and within topics. In addition, the validity of the proposed method was verified using a standard Chinese news corpus. In terms of practical application, our evaluation method was also feasible in the case of public opinion analysis during the COVID-19 epidemic period. The temporal and spatial distributions of five topics generated by the LDA model were all closely related to the development trends of the COVID-19 epidemic. As the results of LDA depend closely on the text content, it is necessary to complete corpus cleaning and filtering carefully when selecting topic numbers based on this method.

(2) Considering the spatial and temporal distribution characteristics of check-in microblogs in Wuhan, there was a certain correlation between check-in microblogs and the number of confirmed COVID-19 cases. Therefore, it is still feasible to analyze public opinion using the text content of check-in microblogs, instead of all microblogs, although the results may be biased. In addition, on the basis of obtaining five topic categories after text classification using LDA, the check-in microblogs also contained temporal and spatial information, which can be utilized to further analyze public opinion topics in Wuhan from temporal and spatial aspects.

(3) Our research showed that the variation trends of the time-series of the different topics all peaked on the day when Wuhan city went into lockdown, and the fluctuation characteristics were consistent with the text contents of various topics. For example, "appreciation and praying" reached a peak on the National Day of Mourning. The trends of the topics over time were in sync with the development of the epidemic situation, indicating that the public is greatly influenced by the Internet and policies, while the COVID-19 epidemic was gradually suppressed.

(4) The spatial distributions of the public opinion topics were mainly clustered in the residential areas within the Fifth Ring Road of Wuhan. On one hand, due to the fact that most people stayed at home for the purposes of epidemic prevention and control, there were relatively slight differences among the spatial distributions of the different topics. Among them, a hot-spot of the "appreciation and praying" topic appeared around Leishenshan Hospital. On the other hand, check-in posts on Weibo are highly correlated with the population distribution. The population density in the center of Wuhan was higher than the surrounding areas, such that the urban center remained a hot-spot for check-in behavior. According to the spatial distribution of diverse public opinion topics, various other hot-spots could be obtained. Based on the above analysis results, differentiated management of public opinion can be executed, and the direction of public opinion can be accurately guided.

It should be noted that the BERT model adopted in this paper was only used for text vectorization, and did not improve the original LDA topic model; therefore, it is necessary to improve the LDA model for text classification on Weibo in the future. Secondly, as the home addresses of confirmed COVID-19 patients are private, it is difficult to explore the spatial correlation between public opinion on Weibo and the actual epidemic situation within different communities. A greater focus on this aspect could produce more interesting findings.

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

Funding: This research was funded by National Key R&D Program of China, grant number 2018YFC0809100.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors would like to thank Yong Hu et al. for providing a large-scale social media dataset from Weibo. The authors are grateful to Xiao for his bert-as-service Python module.

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

References

1. The 45th China Statistical Report on Internet Development. 2020; p. 19. Available online: http://www.cac.gov.cn/2020-04/27/c_1589535470378587.htm (accessed on 24 June 2020). (In Chinese)
2. 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. (In Chinese) [CrossRef]
3. Du, Y.; Xu, J.; Zhong, L.; Hou, Y.; Shen, J. Analysis and Visualization of Multi-Dimensional Characteristics of Network Public Opinion Situation and Sentiment: Taking COVID-19 Epidemic as an Example. *J. Geo-Inf. Sci.* **2021**, *23*, 318–330. (In Chinese) [CrossRef]
4. Debnath, R.; Bardhan, R. India Nudges to Contain COVID-19 Pandemic: A Reactive Public Policy Analysis Using Machine-Learning Based Topic Modelling. *PLoS ONE* **2020**, *15*, e0238972. [CrossRef] [PubMed]
5. Zheng, H.; Goh, D.H.-L.; Lee, C.S.; Lee, E.W.J.; Theng, Y.L. Uncovering Temporal Differences in COVID-19 Tweets. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e233. [CrossRef]
6. 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]
7. Kang, Y.; Wang, Y.; Zhang, D.; Zhou, L. The Public's Opinions on a New School Meals Policy for Childhood Obesity Prevention in the U.S.: A Social Media Analytics Approach. *Int. J. Med. Inform.* **2017**, *103*, 83–88. [CrossRef]
8. Wu, J.; Sivaraman, V.; Kumar, D.; Banda, J.M.; Sontag, D. Pulse of the Pandemic: Iterative Topic Filtering for Clinical Information Extraction from Social Media. *J. Biomed. Inform.* **2021**, *120*, 103844. [CrossRef]
9. Gorodnichenko, Y.; Pham, T.; Talavera, O. Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection. *Eur. Econ. Rev.* **2021**, *136*, 103772. [CrossRef]
10. Krasnov, F.; Sen, A. The Number of Topics Optimization: Clustering Approach. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 416–426. [CrossRef]
11. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Barachi, M.E.; 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]
13. 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]
14. Shibuya, Y.; Tanaka, H. Public Sentiment and Demand for Used Cars after a Large-Scale Disaster: Social Media Sentiment Analysis with Facebook Pages 2018. *arXiv* **2018**, arXiv:1801.07004.
15. 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]
16. Zhang, C.; Ma, X.; Zhou, Y.; Guo, R. Analysis of Public Opinion Evolution in COVID-19 Pandemic from a Perspective of Sentiment Variation. *J. Geo-Inf. Sci.* **2021**, *23*, 341–350. (In Chinese) [CrossRef]
17. Chen, X.-S.; Chang, T.-Y.; Wang, H.-Z.; Zhao, Z.-L.; Zhang, J. Spatial and Temporal Analysis on Public Opinion Evolution of Epidemic Situation about Novel Coronavirus Pneumonia Based on Micro-Blog Data. *J. Sichuan Univ.* **2020**, *57*, 409–416. (In Chinese)
18. 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*, e21978. [CrossRef]
19. Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. A Density-Based Method for Adaptive LDA Model Selection. *Neurocomputing* **2009**, *72*, 1775–1781. [CrossRef]
20. Deveaud, R.; Sanjuan, E.; Bellot, P. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Doc. Numér.* **2014**, *17*, 61–84. [CrossRef]
21. Han, K.; Xing, Z.; Liu, Z.; Liu, J.; Zhang, X. Research on Public Opinion Analysis Methods in Major Public Health Events: Take COVID-19 Epidemic as an Example. *J. Geo-Inf. Sci.* **2021**, *23*, 331–340. (In Chinese) [CrossRef]
22. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
23. 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]
24. Wang, Y.; Li, H.; Wang, T.; Zhu, J. The Mining and Analysis of Emergency Information in Sudden Events Based on Social Media. *Geomat. Inf. Sci. Wuhan Univ.* **2016**, *41*, 290–297. (In Chinese) [CrossRef]
25. 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]

26. Guo, J. Classification for Chinese Short Text Based on Multi LDA Models. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2014. (In Chinese).
27. Wang, T.; Han, M.; Wang, Y. Optimizing LDA Model with Various Topic Numbers: Case Study of Scientific Literature. *Data Anal. Knowl. Discov.* **2018**, *2*, 29–40. (In Chinese) [CrossRef]
28. Griffiths, T.; Steyvers, M. Finding Scientific Topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101* (Suppl. 1), 5228–5235. [CrossRef] [PubMed]
29. Arun, R.; Suresh, V.; Madhavan, C.E.V.; Murthy, M.N.N. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery & Data Mining, Hyderabad, India, 21–24 June 2010.
30. Li, L.; Zhao, X. A Research Summary of Topic Discovery Methods Based on Topic Model. *J. MUC* **2021**, *30*, 59–66. (In Chinese)
31. Guan, P.; Wang, Y.; Fu, Z. Effect Analysis of Scientific Literature Topic Extraction Based on LDA Topic Model with Different Corpus. Libr. *Inf. Serv.* **2016**, *60*, 10. (In Chinese) [CrossRef]
32. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc.* **2006**, *101*, 1566–1581. [CrossRef]
33. Ignatenko, V.; Koltcov, S.; Staab, S.; Boukhers, Z. Fractal Approach for Determining the Optimal Number of Topics in the Field of Topic Modeling. *J. Phys. Conf. Ser.* **2019**, *1163*, 012025. [CrossRef]
34. Koltcov, S. Application of Rényi and Tsallis Entropies to Topic Modeling Optimization. *Phys. A Stat. Mech. Its Appl.* **2018**, *512*, 1192–1204. [CrossRef]
35. Chen, E.; Jiang, E. Review of Studies on Text Similarity Measures. *Data Anal. Knowl. Discov.* **2017**, *1*, 1–11. (In Chinese)
36. Ma, C. The Hitchhiker's Guide to LDA. *arXiv* **2019**, arXiv:1908.03142. (In Chinese)
37. Vayansky, I.; Kumar, S.A.P. A Review of Topic Modeling Methods. *Inf. Syst.* **2020**, *94*, 101582. [CrossRef]
38. Smith, H.; Cipolli, W. The Instagram/Facebook Ban on Graphic Self-Harm Imagery: A Sentiment Analysis and Topic Modeling Approach. *Policy Internet* **2021**. [CrossRef]
39. Scikit-Learn: Machine Learning in Python. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition (accessed on 30 March 2021).
40. Kang, Y.; Wang, Y.; Zhang, D.; Zhou, L.; Sun, M.; Li, J.; Guo, Z.; Zhao, Y.; Zheng, Y.; Si, X.; et al. THUCTC: An Efficient Chinese Text Classifier. Available online: http://thuctc.thunlp.org/ (accessed on 30 April 2021).
41. Nikita, M. Ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. Available online: https://CRAN.R-project.org/package=ldatuning (accessed on 30 March 2021).
42. Hu, Y.; Huang, H.; Chen, A.; Mao, X.-L. Weibo-COV: A Large-Scale COVID-19 Social Media Dataset from Weibo. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online, December 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020.
43. Hu, Y.; Huang, H.; Chen, A.; Mao, X.-L. Weibo-Public-Opinion-Datasets. Available online: https://github.com/nghuyong/weibo-public-opinion-datasets (accessed on 24 June 2020).
44. Full Daily Statistics of 2019-NCoV. Available online: https://github.com/canghailan/Wuhan-2019-nCoV (accessed on 1 February 2021).
45. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
46. Weibo User Development Report in 2020. 2021, p. 4. Available online: https://data.weibo.com/report/reportDetail?id=456 (accessed on 24 June 2020). (In Chinese)
47. Xiao, H. Bert-as-Service. Available online: https://github.com/hanxiao/bert-as-service (accessed on 30 April 2021).