


## Article

# A Normalized Rich-Club Connectivity-Based Strategy for Keyword Selection in Social Media Analysis

Ying Lian <sup>1</sup>, Xiaofeng Lin <sup>2</sup> , Xuefan Dong <sup>3,4</sup> and Shengjie Hou <sup>5,\*</sup><sup>1</sup> School of Journalism, Communication University of China, Beijing 100024, China; lianying@cuc.edu.cn<sup>2</sup> Center for Innovation-Driven Development (Center for Digital Economy Research and Development), National Development and Reform Commission, Beijing 100045, China; linxfcxzx@163.com<sup>3</sup> College of Economics and Management, Beijing University of Technology, Beijing 100124, China; dongxf@bjut.edu.cn<sup>4</sup> Research Base of Beijing Modern Manufacturing Development, Beijing University of Technology, Beijing 100124, China<sup>5</sup> National Innovation Institute of Defense Technology, Beijing 100071, China

\* Correspondence: houshengjiework@sina.com

**Abstract:** In this paper, we present a study on keyword selection behavior in social media analysis that is focused on particular topics, and propose a new effective strategy that considers the co-occurrence relationships between keywords and uses graph-based techniques. In particular, we used the normalized rich-club connectivity considering the weighted degree, closeness centrality, betweenness centrality and PageRank values to measure a subgroup of highly connected “rich keywords” in a keyword co-occurrence network. Community detection is subsequently applied to identify several keyword combinations that are able to accurately and comprehensively represent the researched topic. The empirical results based on four topics and comparing four existing models confirm the performance of our proposed strategy in promoting the quantity and ensuring the quality of data related to particular topics collected from social media. Overall, our findings are expected to offer useful guidelines on how to select keywords for social media-based studies and thus further increase the reliability and validity of their respective conclusions.

**Keywords:** keyword selection; social media; co-occurrence relationship; rich-club



**Citation:** Lian, Y.; Lin, X.; Dong, X.; Hou, S. A Normalized Rich-Club Connectivity-Based Strategy for Keyword Selection in Social Media Analysis. *Sustainability* **2022**, *14*, 7722. <https://doi.org/10.3390/su14137722>

Academic Editor: Andreas Kanavos

Received: 7 May 2022

Accepted: 16 June 2022

Published: 24 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Keywords can be defined as special words embedded in a document that can provide a precise and accurate representation of that document's content [1]. In many areas, keyword selection is a primary step, such as in text mining [1,2], bibliometrics [3–5] and communication [6,7]. In general, selecting the appropriate keywords is helpful for ensuring the quality of collected data, reducing the costs for data cleaning and accordingly increasing the rigor and significance of findings.

In recent years, social media, such as Twitter, Facebook and Weibo, has been cited as an important tool for scholars investigating a variety of fields, including social sciences [8], communication [6,7], politics [9,10], education [11,12], medical science [13], transportation [14] and disaster management [15,16]. As compared to traditional data collection channels, such as surveys or questionnaires, social media can yield a better understanding of the public perceptions, decrease time and commercial costs in the data collection process, and display a greater variety and geographic distribution in the data [17–19]. However, since the data size in social media is generally very large, redundant and invalid information should be controlled [20], such as rumors, advertising, purposeful information posted by spammers, information with a disordered format or missing content, forwarding information and information with too few words or that makes no actual sense. These data may have a negative influence on conclusions and accordingly affect the correctness and

effectiveness of decisions. In order to solve this problem, the data cleaning process plays an important role in social media analysis [17,21,22]. However, data cleaning is time consuming for social media-based data and there is currently no way to examine accuracy that is both rapid and effective. Therefore, using a set of appropriate keywords that can accurately and comprehensively represent the researched topic is a practical alternative. Regarding the keywords in social media analysis, the primary issue that should be considered is how to determine the list of keywords in order to obtain the maximum number of posts that are closely related to the topic. Specifically, an effective keyword list can also ensure the quality of the collected data and reduce the time costs of the subsequent data cleaning process [23,24]. On the other hand, if the selected keywords cannot adequately represent the characteristics of the entire topic, the reliability and accuracy of the subsequent analysis may be affected, which will in turn make it difficult to obtain reliable insights from the results [3].

However, previous studies have commonly neglected this process by assuming that the keywords used in their studies were extracted well and there was no need to examine the applicability of the keywords. For example, Han and Wang [25] studied the online posts on Weibo regarding the flood that occurred in Shouguang, a county-level city in Shandong Province, China. However, the only keyword they used was “Shouguang”, with no further elaboration beyond this single term. This may be due to the fact that the majority of online posts in social media containing the word “Shouguang” were related to the flood within their study’s time interval. Nevertheless, there can be no doubt that using “Shouguang” as the only keyword will result in a large amount of irrelevant data that could thus negatively influence their final conclusions. In this case, we believe that the addition of “flood” to the keyword list would have been helpful.

In most social media analyses, there are some problems in the keyword selection process. For instance, because of multiple meanings for the same word in different contexts, irrelevant data can be collected. Cody et al. [26] used “climate” as a keyword when collecting public opinions about climate change; however, they found that not every tweet collected was about climate change. Thus, they had to filter the data manually, which consumes a considerable amount of time. Furthermore, Noh et al. [2] stated that the method for determining the number of keywords to select is critical, as it will affect the quantity and quality of the collected data. Specifically, if the number of keywords is relatively large, the quality will improve while the quantity decreases; if it is too few, the quantity will increase while the quality decreases.

Taking into consideration the current state of the literature, to our knowledge, few studies to date have been conducted on understanding the role of the keyword in social media analysis. Wang et al. [27] carried out a pioneer study by putting forward a novel technique named Double Ranking (DR), in which there are two steps: the first is to provide some general keywords related to the research topic according to the personal experience, and the second step is to use these words to collect possibly relevant online posts and extract more keywords based on the results of two rounds of rankings. This work firstly highlighted the importance of keyword identification in social media analysis and offered valuable explanations about the differences between it and traditional keyword extraction problem. As an extended study, Zheng and Sun [28] proposed an automatic keyword generation model by applying the machine learning technique and three properties: relevance, coverage and evolvement. However, this model could be only effectively applied for topics about major events, in which the number of online posts within each time window is substantial. In addition, there is no appropriate method to set the size of time window, especially for topics with short durations. Thus, putting forward a novel keyword selection method with a wider applied range of topics for social media is of great significance. In order to address this gap in the literature, in the current study, we provide an investigation of keyword selection in social media analysis and propose a new keyword selection strategy based on the social network analysis.

The original contributions of our research are listed below.

1. This paper highlights the important role played by the keyword selection process in social media analysis, a topic that has generally been neglected by most prior studies. We claim that the use of the appropriate keywords for data collection can improve the quality of the data and accordingly greatly enhance the significance of a study. Thus, this paper contributes to enhancing researchers' attention to keyword selection in the process of using social media data for analyzing, which could yield more accurate and persuasive research results, to a large extent.
2. Using a graph-based approach, we propose a new keyword selection method for social media analysis considering two different types of topics: conceptual topics and event-based topics. In particular, the normalized rich-club connectivity considering the weighted degree, closeness centrality, betweenness centrality and PageRank values are used to identify "rich keywords", and community detection is applied to determine the keyword combinations for representing the research topic.
3. We evaluate our method by using the data related to four topics and comparing with four widely used keyword selection techniques. According to the results of the empirical test, our method can reach a balance between the quantity and quality of the data. In other words, it can greatly increase the amount of high-quality data. In addition, since social media is an essential data source for a variety of areas, especially for those studying public perceptions, our proposed keyword selection method can benefit future studies in various areas, including decision making, disaster management and policy development.

The overall structure of this paper is as follows. Section 2 describes issues with keywords in social media analysis, including the general framework of social media analysis, kinds of topics studied and existing keyword selection methods. The newly proposed keyword selection strategy is described in Sections 3 and 4, providing a comparison analysis to test the performance of the proposed strategy. The research results are described and discussed in Section 5, and finally, Section 6 provides the conclusion and some directions for future research.

## 2. Keywords in Social Media Analysis

### 2.1. Framework for Social Media Analysis

A framework for social media analysis has been built by existing studies, which includes four main steps: topic determination, data collection, data preprocessing and data analysis [29]. The first step in social media analysis is to determine the topic. For this step, the targeted topic should be of concern to a large number of users on social media in order to guarantee a sufficient quantity of available data. For the second step, there are three primary elements: the data source, keywords and time interval. Specifically, the data source refers to the selected social media to be used for the data collection, such as Twitter, Weibo or Facebook, and the time interval includes the determination of the start and end points to be considered in the data collection [30]. The generally used method for determining the time interval is to identify the number of related messages day by day or hour by hour, setting the end point as the time when a minimum number of messages or no messages appear [31]. The issues regarding keywords are demonstrated in Sections 2.2 and 2.3. In addition, data preprocessing is an essential step to ensure the accuracy and effectiveness of the data. First, the collected data must be cleaned, which mainly focuses on the text of the online posts. Data with a disordered format, missing content, too few words or no actual meaning should be removed in order to reduce their negative influence on the findings. Then, since the initially collected datasets from social media are generally unstructured [23,24,32], they should be transformed into a structured form in order to identify the basic elements contained in the data.

There are four main parts to the data analysis: temporal and spatial distribution analysis, user analysis, topic identification and sentiment analysis. The first part consists of the simple statistics regarding the number of posts from temporal and spatial perspectives [11,29]. The second part mainly focuses on the formed communication network

structure [33] and personal characteristics of the users [34], such as gender, occupation and location. This part of the analysis can provide a detailed understanding of the different populations that are paying more attention to a particular topic and how information spreads on social media. The third part identifies the main viewpoints found in the data, for which text clustering algorithms are commonly applied [35]. Finally, the aim of the fourth part seeks to represent the sentiment patterns and their changes over time in the collected data [36–38]. In sum, the quantity and quality of the collected data are the basis of ensuring the accuracy and significance of the empirical results as well as any subsequently obtained conclusions. For instance, if many online posts related to the research topic are missed, the temporal and spatial analysis may not provide an accurate picture reflecting the real situation of the topic. In addition, if some unrelated posts are contained in the analyzed dataset, the findings for sentiment identification and viewpoint clustering may lead to an incorrect understanding of public opinions [39].

## 2.2. Types of Topics

Different kinds of keywords should be used for different kinds of targeted topics. Based on existing studies, we identified two kinds of topics, namely, conceptual topics and event-based topics, as shown as Table 1. The first category generally appears in studies focused on a particular concept [40], for example, green buildings and climate change. The second category is frequently applied in studies focusing on a specific event [28], such as Hurricane Harvey and the Shouguang city flood.

**Table 1.** Categories of topics in social media analysis.

Pattern	Name	Examples
Pattern 1	Conceptual topics	Green buildings [22], transportation planning [14], climate change [26], low-carbon city [41]
Pattern 2	Event-based topics	The Xiangtan Pregnant Woman Event [42], “I will never go to Hong Kong again!” [21], Hurricane Harvey [33], the Shouguang City Flood [25]

## 2.3. Existing Keyword Selection Methods

Based on previous studies, we identified four existing keyword selection methods for social media analysis. For the first method, as most studies have not presented details on their keyword selection process [22,25,41], we summarized this kind of keyword selection method as the experience-based (EB) method. The second method is the snowball-sampling (SS) method [23]. In this method, scholars first collect a small number of messages using a keyword that is directly related to the topic and subsequently identify the most frequently co-occurring keywords. Then, the collection process is repeated with the additional identified keywords. The third method is the DR model proposed by Wang et al. [27], which uses a two-step ranking mechanism to select keywords. The final method is the topic modeling algorithm, such as Latent Dirichlet allocation (LDA) [43] and Targeted Topic Modeling (TTM) [44]. In this kind of method, keywords are extracted and clustered based on the similarity metrics to obtain latent topics.

In addition, Zheng and Sun [28] put forward the ALMIK (Active Learning based on Multiple-Instance learning with Keyword extraction) model to identify potential useful posts from a large set of online posts during disasters by using the Convolutional Neural Networks (CNN) machine learning algorithm. However, this model is outside the range of our study, as it is used for classifying the collected data rather than collecting potentially useful data from social media.

## 2.4. Network Techniques in Keyword Selection

Identifying semantically related terms based on the co-occurrence correlation is a research topic with a long tradition in linguistics and information theory [5,45]. With

respect to studies using the co-occurrence network of words for topic text extraction, community detection algorithms, such as fast-unfolding algorithm [46], are generally employed to cluster the keywords contained in a given document. Community detection can be defined as the identification of the community structure of a network according to its topological relationships [47]. In most cases, relatively strong connections can be discovered between the nodes contained in a similar community. Wagenseller et al. [48] carried out a comparative analysis focusing on different existing community detection algorithms and put forward some useful future directions in this area. In the present paper, we employed network-based techniques to propose a new keyword selection method for social media analysis.

### 3. Graph-Based Keyword Selection Strategy

#### 3.1. Network Construction

We modeled the texts of online posts as a weighted, undirected network:  $G = (V, E, W)$ , where  $V$  is the set of vertices that represents the keywords,  $E \in V \times V$  is the set of edges that measures the co-occurrence relationship between each pair of nodes and  $W$  is the corresponding weighted adjacency matrix. Two keywords are linked only if they co-occur in a similar online post, and the weight between them refers to their number of occurrences [45].

#### 3.2. Rich-Club Phenomenon

##### 3.2.1. Concept

The rich-club phenomenon is defined as a particular phenomenon in networks in which subgroups of important or influential (rich) nodes preferentially and intensely interact with one another [49]. As this phenomenon focuses on both the inner cycle among rich nodes and the links between rich and non-rich nodes [50], it has been applied in a variety of areas to explore the characteristics of real word networks, such as the human brain [51], population flow [52] and patient referral behavior [50].

In this paper, the rich-club phenomenon is taken into consideration because it is helpful for increasing the generality and relevance of keywords to the researched topics when it is used in the selection of keywords for social media-based studies. In a keyword co-occurrence network for a particular topic, an identified rich-club phenomenon suggests the existence of a highly connected group of “rich keywords”, which are the most suitable keywords for representing the topic. Furthermore, as there are often multiple meanings for the same word in different contexts, we selected some additional keywords related to the rich keywords to form several keyword combinations through the use of the community detection approach. The detection of the rich-club phenomena provides a new method for keyword selection for social media-based studies that enables a much easier selection of keywords that can more accurately and comprehensively represent a topic and provide more reasonable quantitative evidence. In particular, topics with keyword co-occurrence networks that contain the rich-club phenomena indicate the presence of several centralized and focused sub-topics, while for those without the rich-club effect, there may be more diverse sub-topics.

Moreover, it is noteworthy that while the use of graph techniques has contributed to the keyword selection process in social media-based studies, we still highlight the essential and indispensable role played by personal experience. Our proposed quantitative strategy should only be considered as an assistance tool for reducing time costs and discovering more appropriate keywords and not as a substitute for personal experience. In other words, the results of each step need to be examined and adjusted manually.

##### 3.2.2. Evaluation Metrics

Previous studies regarding the rich-club phenomenon have provided notable contributions to network sciences and proposed a variety of frameworks with different structural characteristics for examining this phenomenon [50,52,53]. In this paper, we applied rich-

club connectivity to identify the “rich keywords” from the co-occurrence networks, in which the weighted degree, closeness centrality, betweenness centrality and PageRank values were considered. Community detection was used to select additional keywords based on the results of “rich keywords” determination.

a. Rich-club connectivity

According to Ren et al. [54], rich-club connectivity is defined as the ratio between the number of existing edges among the nodes in a rich club and the number of all possible edges among these. The mathematical expression is shown in Formula (1):

$$\delta(\xi) = \frac{L_{>\xi}}{\xi(\xi - 1)}, \quad (1)$$

where  $\delta(\xi)$  is the rich-club connectivity and  $\xi$  is a threshold parameter used to determine the rich node. Specifically, a node with a node strength larger than  $\xi$  is considered as a rich node.  $L_{>\xi}$  represents the number of edges among the rich nodes. In particular,  $\delta(\xi) = 1$  means that all members of the rich club are connected with each other and  $\delta(\xi) = 0$  indicates that there are no connections between any two members of the rich club. Furthermore, to evaluate the statistical significance of  $\delta(\xi)$ , this coefficient is typically normalized by using building a random network with preserved degree distribution [55]. The mathematical expression is shown in Formula (2):

$$\delta_{norm}(\xi) = \frac{\delta(\xi)}{\langle \delta_{random}(\xi) \rangle}, \quad (2)$$

where  $\langle \rangle$  represents the average and  $\delta_{norm}(\xi)$  represents the normalized rich-club connectivity, which is a ratio between the rich-club connectivity of the studied network and that of a randomized network.  $\delta_{random}(\xi)$  measures the rich-club connectivity of the randomized network. According to Tang et al. [50], a positive rich-club effect can be confirmed if  $\delta_{norm}(\xi) > 1$ . There are two ways to determine the value of  $\xi$ , namely by degree or weighted degree.

Cinelli [56] proposed a generalized rich-club framework for computing the normalized rich-club connectivity in terms of other structural measures distinct to degree, displayed as Formula (3):

$$\delta_{norm}(\alpha) = \frac{\delta(\alpha)}{\langle \delta_{random}(\alpha) \rangle}, \quad (3)$$

where the value of  $\alpha$  corresponds to structural measures of the network. In this paper, by using Formula (3), in order to comprehensively consider the node importance in rich club identification, the weighted degree, closeness centrality, betweenness centrality and PageRank values were all used to calculate rich-club connectivity. Thus, we have four rich-club connectivity metrics, which were measured by  $\delta_{norm}(W)$ ,  $\delta_{norm}(C)$ ,  $\delta_{norm}(B)$  and  $\delta_{norm}(P)$ . In addition, as the weighted degree, closeness centrality, betweenness centrality and PageRank values have been commonly applied by previous studies [57–59] for analyzing particular networks, details about their calculation methods are not provided in this paper.

In terms of the formation rule of randomized network, based on Maslov and Snep-pen [60], we repeated the following procedure until the weight of every edge had been changed: two edges contained in the studied network were randomly selected and their weights were automatically redistributed by remaining attached to the reshuffled edges subsequently. In addition, if either of these edges was already formed, two other new edges were selected. According to previous studies [61,62], this rule could offer a higher degree of randomization and ensure that the weight distribution and degree distribution remain unchanged.

Moreover, according to previous studies [63], the demarcation points that appeared exactly before significant decrease and fluctuations in the rich-club connectivity curve were

commonly applied to determine the rich nodes. However, this method is based on personal observations rather than numerical thresholds, which may yield confusion results. Thus, the volatility rate was employed in this paper. Its mathematical expression is shown in Formula (4):

$$R = \frac{\delta_{norm}(\alpha)_t}{\delta_{norm}(\alpha)_{t-1}} - 1 \quad (4)$$

where  $R_t$  is the volatility rate and  $t$  measures the computation sequence. In particular, the point with the first relatively large negative  $R_t$  value is set as the demarcation point to determine the “rich keywords”. It should be noted that the volatility rate is an auxiliary reference index and artificial judgement is still needed in this process.

#### b. Community detection

We employed the fast-unfolding algorithm to perform community detection. In particular, the fast-unfolding algorithm is used as the basis of a modularity function, ranging from  $-1$  to  $1$  [64]. The largest value for modularity indicates the optimum community classification results. Formula (5) shows the computation of the modularity:

$$Q = \frac{1}{2m} \sum_{ij} \left[ e_{ij} - \frac{\lambda_i \lambda_j}{2m} \right] \omega(c_i, c_j), \quad (5)$$

where  $Q$  is the modularity,  $e_{ij}$  measures the weight of the edge linking nodes  $i$  and  $j$ ;  $\lambda_i$  and  $\lambda_j$  measure the sum of the weights of the edges linking to nodes  $i$  and  $j$ , respectively;  $c_i$  represents the community to which node  $i$  is assigned;  $\delta$  is a simple delta function and  $\omega(c_i, c_j) = 1$  if  $c_i = c_j$ ;  $\omega(c_i, c_j) = 0$  if  $c_i \neq c_j$ ; and  $2m$  is the sum of the weight of all edges. For keyword co-occurrence networks, community detection can show the strength of the co-occurrence relationship between different keywords.

### 3.3. Keyword Selection Strategy

In this section, we propose a new graph-based keyword selection strategy for social media analysis consisting of the following steps (displayed in Figure 1).

**Step 1:** In this step, we need to consider the aforementioned different patterns of topics.

- **Pattern 1 topics:** Set the primary keywords for the topic based on the research target. The number of primary keywords should be as small as possible because a larger number will limit the size of the data.
- **Pattern 2 topics:** Using the initial text of the topic, which refers to the first posted information describing the studied policy or event, perform text segmentation and identify no more than three primary keywords based on personal experience and word frequency. The reason for not using graph-based techniques in this step is that the length of the initial text of the topic is usually short and will result in a highly dense network [1]. In other words, considering the co-occurrence relationship between keywords during this step is mostly meaningless.

**Step 2:** Use the extracted primary keywords and determined time interval to collect related online posts from social media. It should be noted that there is no need to conduct the data cleaning during this step because the utilization of the extracted primary keywords can ensure that the majority of the collected online posts are about the research topic. Furthermore, the negative influences caused by other unrelated posts can be eliminated by using graph-based techniques during the following steps.

**Step 3:** Perform text segmentation based on the collected online posts and accordingly build the keyword co-occurrence network.

**Step 4:** Build 50 randomized networks based on the rule proposed in Section 3.2.1 and calculate the values of  $\delta_{norm}(W)$ ,  $\delta_{norm}(C)$ ,  $\delta_{norm}(B)$ ,  $\delta_{norm}(P)$  and  $R_t$  according to Formulas (1), (3) and (4). Using  $\delta_{norm}(W)$ ,  $\delta_{norm}(C)$ ,  $\delta_{norm}(B)$ ,  $\delta_{norm}(P)$  and the volatility rate  $R_t$  to determine “rich keywords” from the constructed co-occurrence network, in which only those identified by all four measures are selected.

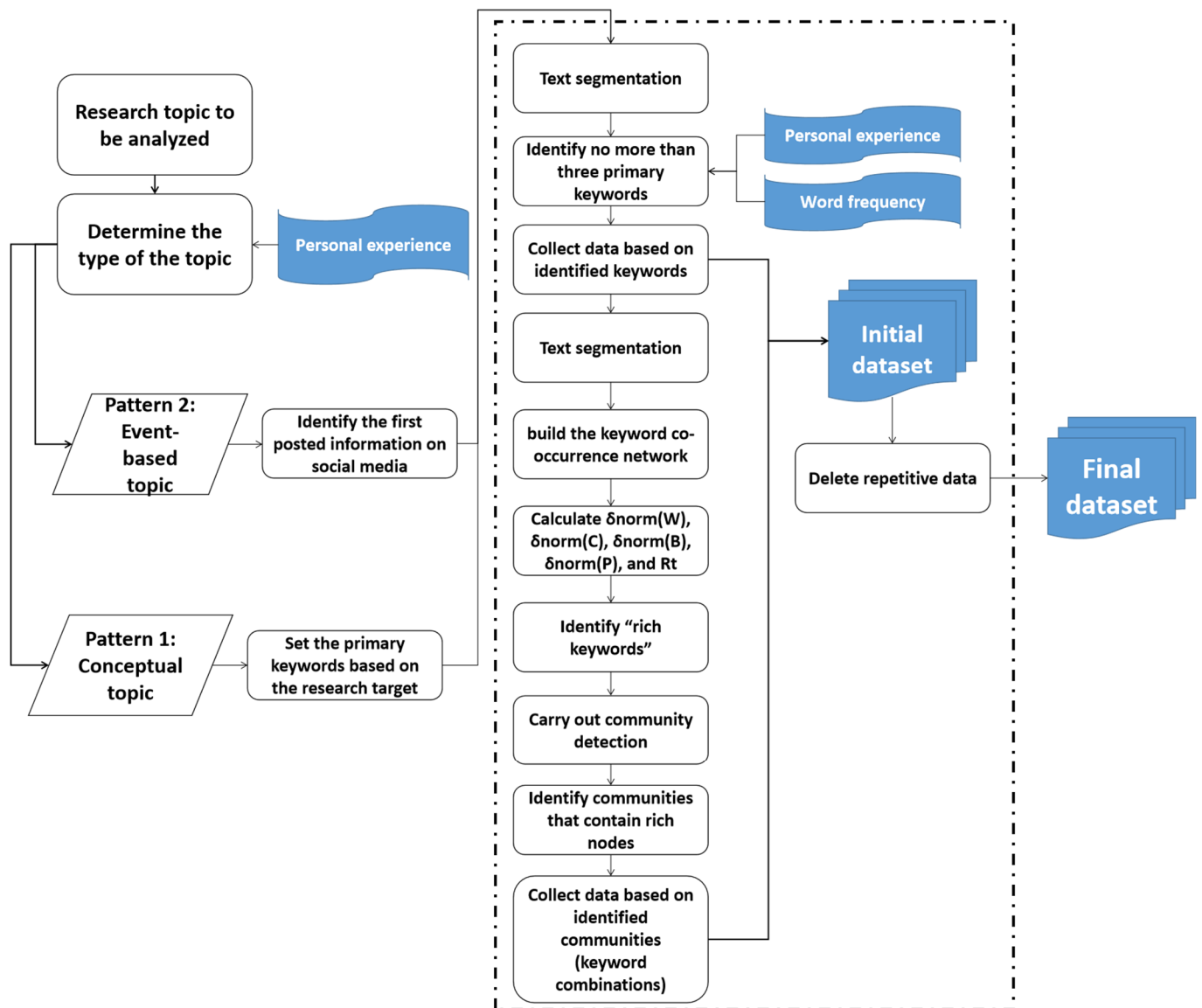


Figure 1. Flowchart of proposed keyword selection method.

**Step 5:** Perform community detection on the built keyword co-occurrence network and consider only the communities that contain rich nodes. Each selected community refers to a keyword combination that contains only the most connected keywords.

**Step 6:** Use keyword combinations to collect related online posts from social media and subsequently delete the repetitive online posts.

We employed the Jieba Natural Language Toolkit for Python for the text segmentation of collected online posts. Stop words and non-content bearing words, such as prepositions and conjunctions, were deleted by using the stop-word list and the part-of-speech tagging technique, respectively, embedded in the toolkit. In addition, Octopus, a mature web crawler tool, was used to collect the online posts from social media, as it has shown great effectiveness in prior studies [22,29,40]. Information about how to use Octopus can be found on its official website: <https://www.bazhuayu.com/> (accessed on 26 August 2021).

## 4. Empirical Study

### 4.1. Data

In this study, the data source used was Weibo, which is the biggest microblogging site in China. According to the 45th China Statistical Report on Internet Development



published by the China Internet Network Information Center (CNNIC), the number of active Weibo users was 70 million at the end of March 2020, indicating that it attracts more than half of all Chinese netizens on average each month. Additionally, since many government departments, news media providers, enterprises, organizations and influential individuals have accounts with Weibo, it has attracted the attention of a great number of studies [21,22,25,29,65].

We selected four topics for examining the performance of our proposed keyword selection strategy, of which two were of Pattern 1 and two were of Pattern 2. The details of each topic are shown in Table 2. The reasons for choosing these four cases were: (1) the number of related online posts differed between them, (2) they contained different types of topics and (3) the selected topics were relatively new and could avoid the influence on the data caused by time. In sum, the utilization of these four cases could provide a more reasonable and comprehensive experiment for testing the performance of our proposed strategy. In addition, in order to ensure the consistency of the data collection process and to minimize the influences aroused by other factors on data quality, similar data collection software and steps were used in all cases.

**Table 2.** Details of the four selected topics.

	Description	Pattern
Topic 1	Medical violence (called “Yi Nao” in Chinese)	1
Topic 2	Cyber violence	1
Topic 3	National People’s Congress (NPC) and Chinese People’s Political Consultative Conference (CPPCC) Annual Sessions 2020	2
Topic 4	New Permanent Residence Law for Foreigners (Draft Version)	2

## 4.2. Methods

### 4.2.1. Comparison Methods

In this paper, we employed the EB, SS, DR [27] and LDA methods as comparisons. In particular, for the EB method and the seed keywords of DR method, three scholars with research interests in social media analysis were asked to choose a keyword for each topic based on their personal experience and their collectively chosen keywords were selected. For the SS method, there were two steps. In the first step, three scholars were asked to select a small number of keywords directly related to the topic, and in the second step, some additional keywords with high word frequency were selected from online posts collected based on the keywords extracted in the first step. In addition, in terms of the LDA method, we used it to replace our proposed network technique for selecting additional keywords, and the value of  $K$  was determined heuristically.

### 4.2.2. Evaluation Methods

In general, when considering issues about keyword selection, both the precision and recall should be employed to comprehensively evaluate the performance of the models. However, due to the nature of our problem to put forward a novel keyword selection method that could be used for collecting as much useful online posts from social media as possible, we could not obtain the entire data related to a given topic in social media. In other words, the ratio of recall cannot be calculated. Thus, to carry out a relatively overall assessment experiment, by following and extending Wang et al. [27], the number of related online posts  $N_c$  and the proportion of related online posts to the total number of collected online posts  $r$  were both employed the results. In particular, the relevance coefficient was calculated using Formula (6):

$$r = \frac{N_c}{N_T} \times 100\%, \quad (6)$$

where  $r$  is the relevance coefficient,  $N_c$  represents the number of related online posts and  $N_T$  is the total number. In addition, we believe that manual detection is the most accurate approach for determining whether a particular online post is related to a topic. Thus, we employed 10 people who had performed social media-based studies in the past to identify the related online posts manually from the collected data, and those confirmed by more than half of the group were set as the related data.

## 5. Results

In this section, we present the results of our experiments. Table 3 shows the selected time intervals and primary keywords for the four topics. For Topic 1 and Topic 2, the primary keywords were determined based on the personal experience of scholars. For Topic 3 and Topic 4, the primary keywords were set based on their related texts. Then, by using these primary keywords, the online posts from within the set time interval that were related to the four topics were collected from Weibo. The number of collected messages for the four topics was 4201, 8226, 8319 and 5404, respectively. These data were also the results of the EB approach.

**Table 3.** Determined time interval and primary keywords for the four topics.

	Time Interval	Primary Keywords
Topic 1	1 January 2018–31 December 2019	Yi Nao, medical dispute, medical order
Topic 2	1 January 2018–31 December 2019	cyber violence
Topic 3 (Text source of Topic 3: <a href="http://www.ccps.gov.cn/xtt/202005/t20200530_141283.shtml">http://www.ccps.gov.cn/xtt/202005/t20200530_141283.shtml</a> ) (accessed on 26 August 2021)	21 May 2020–28 May 2020	two sessions, NPC, CPPCC
Topic 4 (Text source of Topic 4: <a href="http://www.moj.gov.cn/news/content/2020-02/27/zlk_3242559.html">http://www.moj.gov.cn/news/content/2020-02/27/zlk_3242559.html</a> ) (accessed on 26 August 2021)	27 February 2020–4 March 2020	permanent residence, foreigner

Word segmentation was performed for the collected online posts and the stop words and non-content bearing words were removed from the results. For the SS approach, some additional keywords were selected based on word frequency and personal experience (Table 4). For DR, only two iterations were considered. With respect to LDA, each post was regarded as a mixture of latent topics, and a topic refers to a multinomial distribution over words. In addition, by following related studies [66,67], the trial-and-error method was carried out to determine the optimum value of  $K$ . The values of  $K$  were 5, 5, 6 and 4 for the four topics. Then, the top two ranked keywords of each cluster were selected as a keyword combination. Table 4 shows the additional keywords for each topic. Finally, for SS, DR and LDA, the online posts related to each topic were collected from Weibo using the gained additional keywords and the primary keywords, and the duplicate information was removed. The results are shown in Table 4.

Then, for our proposed strategy, keyword co-occurrence networks for each topic were built based on 300 keywords with high word frequency and practical significance, as displayed in Figure 2. In addition to the words that are directly related to four researched topics, such as workplace violence, cyber violence, two sessions and permanent residence, those with higher frequencies mainly include hospital, doctor, fair and patients for Topic 1; network order, flesh search, real name system and keyboard man for Topic 2; Hong Kong, vaccines, epidemic situation and poverty for Topic 3; and racial discrimination, enjoy, justice and international for Topic 4. These results can roughly reflect how the public viewed the researched topics. Then, we calculated the values of  $\delta_{norm}(W)$ ,  $\delta_{norm}(C)$ ,  $\delta_{norm}(B)$  and  $\delta_{norm}(P)$  of four topics' keyword co-occurrence networks by using Formulas (1) and (3), and the results are shown in Figures 3–6, respectively. It can be noticed that the calculated

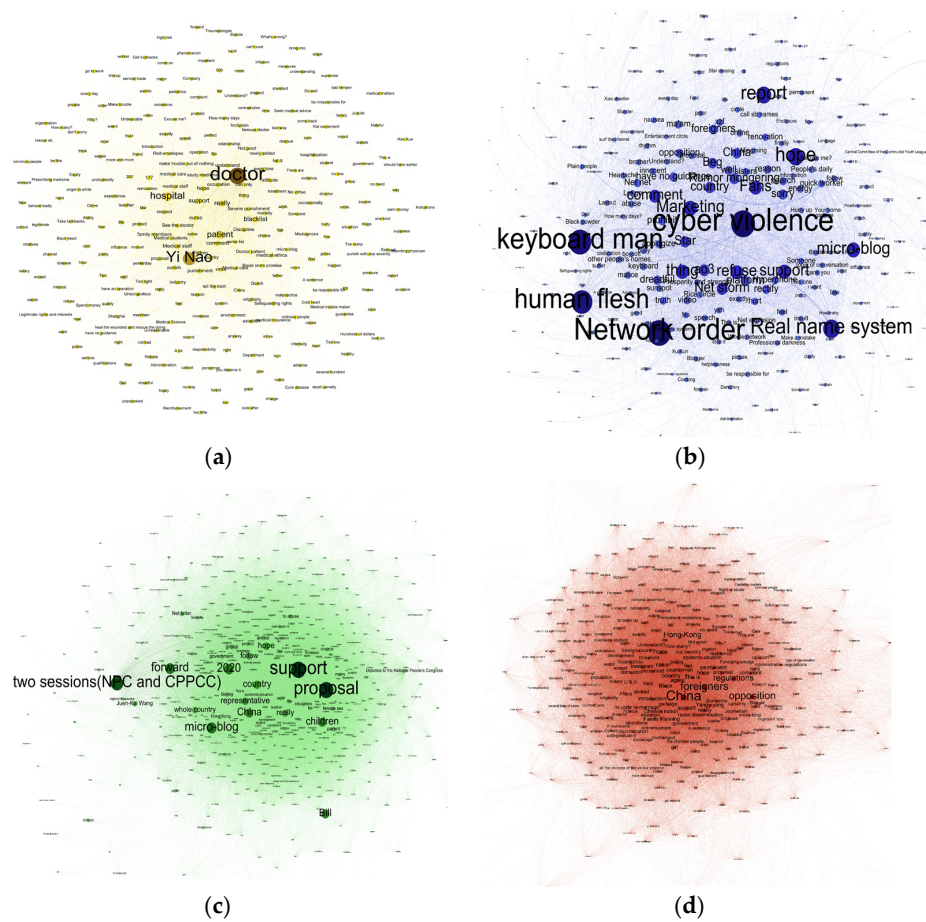
normalized rich-club connectivity values with different  $\xi$  regarding the four topics are all basically greater than 1 and show a general upward trend during the beginning stage of the curves. This finding is consistent with that of Wei et al. [52], indicating a significant rich-club characteristic. In addition, some figures show discrete fluctuations, such as Figure 2a,b, while some others show continuous fluctuation, such as Figure 5b,d. This difference is mainly due to the different topology structure of the four keyword co-occurrence networks.

**Table 4.** Results for the four topics based on the SS, DR and LDA.

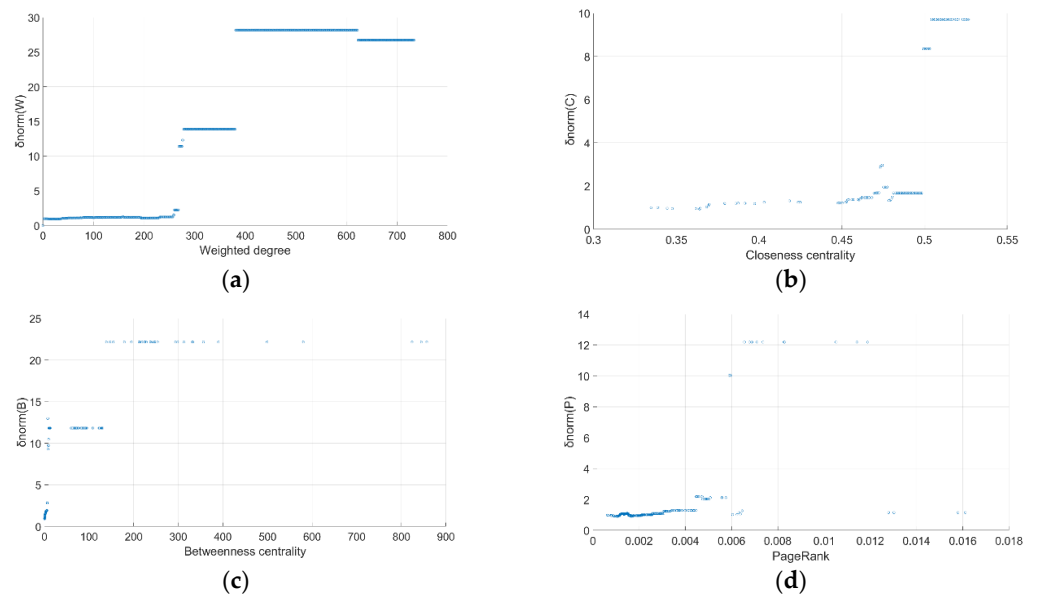
		Additional Keywords	Number of Online Posts in Total
SS	Topic 1	patient right, physician safety, hospital safety, doctor moral	7771
	Topic 2	cyber manhunt, language violence, keyboard man (In China, “keyboard man” refers to a member of a group of individuals who “deliver justice” in the form of comments on the Internet, which they consider “freedom of speech”. However, they are often considered as the main force in launching cyber violence.)	13,592
	Topic 3	civil code, poverty alleviation	10,136
	Topic 4	Ministry of Justice, international students	7930
DR	Topic 1	hospital, blacklist, patient, rebate, doctor, protection	8904
	Topic 2	Weibo, withdraw, law, vicious, free, right, terrible, language, rumor, murder, spreading, supervise	16,527
	Topic 3	poverty, epidemic, civil code, proposal, vaccine, society, Hong Kong	11,074
	Topic 4	residence, citizen, foreign, student, public, right, international, fair	10,026
LDA	Topic 1	doctor + patient, hospital + damage, right + protection, medical staff-law, education + student	9207
	Topic 2	violence + terrible, rumor + video, Weibo + law, Internet + safeguard, statement + response	18,248
	Topic 3	international + cooperation, civil code + protection, well-off + society, corruption + against, Weibo + proposal, hukou + reformation	13,256
	Topic 4	justice + public, foreign + domestic, international + students, residence + fair	8359

Then, values of the volatility rate regarding each normalized rich-club connectivity of four topics were calculated. By following our proposed rules for the “rich keywords” identification, 5, 4, 6 and 3 “rich keywords” were obtained for Topic1, Topic 2, Topic 3 and Topic 4, respectively. Specifically, for Topic 1, the rich keywords are “Yi Nao”, “physician”, “blacklist”, “rights” and “hospital”; for Topic 2, the rich keywords are “cyber violence”, “keyboard man”, “human flesh” and “network order”; for Topic 3, the rich keywords are “two sessions”, “Hong Kong”, “disease”, “proposal”, “poverty” and “vaccine”; for Topic 4, the rich keywords are “Ministry of Justice”, “international students” and “national treatment”.

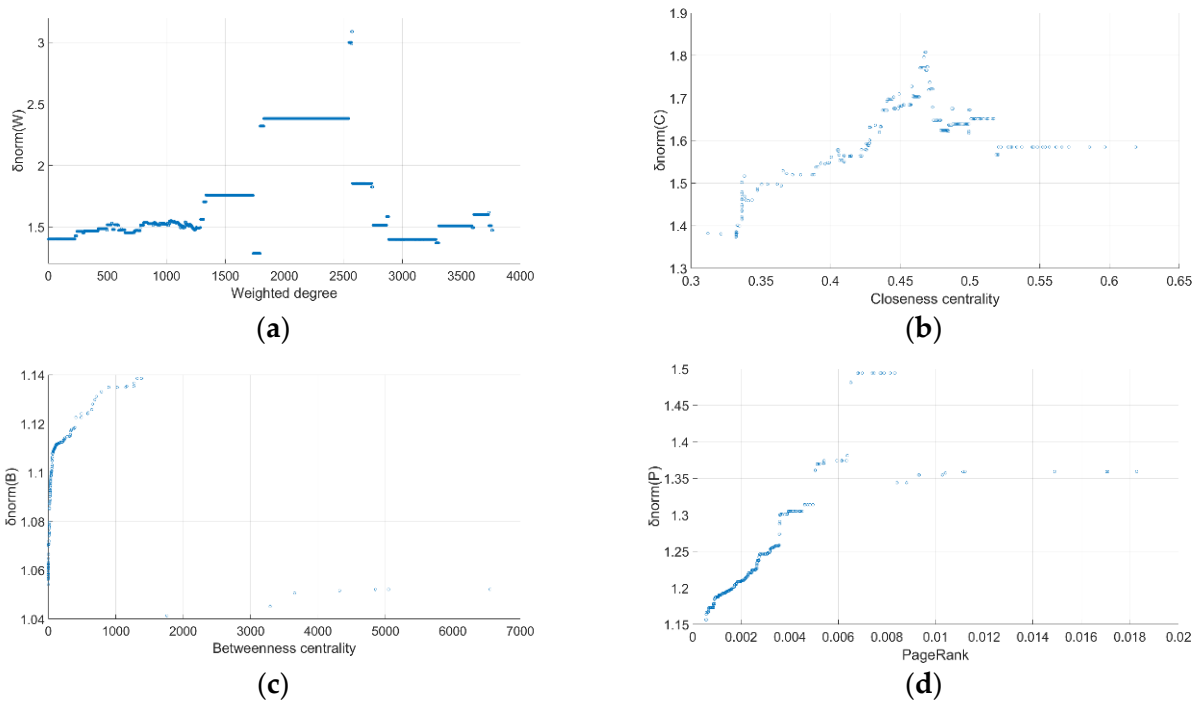
Finally, community detection was performed for all four keyword co-occurrence networks. It should be noted that only the communities containing rich keywords were considered. For example, in Topic 4, there was a community containing the keywords “Weibo”, “pictures”, “comment” and others that could be easily determined as unrelated to the community. Thus, such communities were not considered in the following steps. Then, based on the adjusted community detection results, several additional keyword combinations were formed. For example, for Topic 3, the additional keyword combinations included “Hong Kong + independence”, “disease + control”, “proposal + NPC”, “proposal + CPPCC”, “poverty + alleviation” and “vaccine + HPV”. Then, related online posts for each topic were collected from Weibo using these additional keyword combinations, and the primary keywords and duplicates were removed. The results are shown in Table 5.



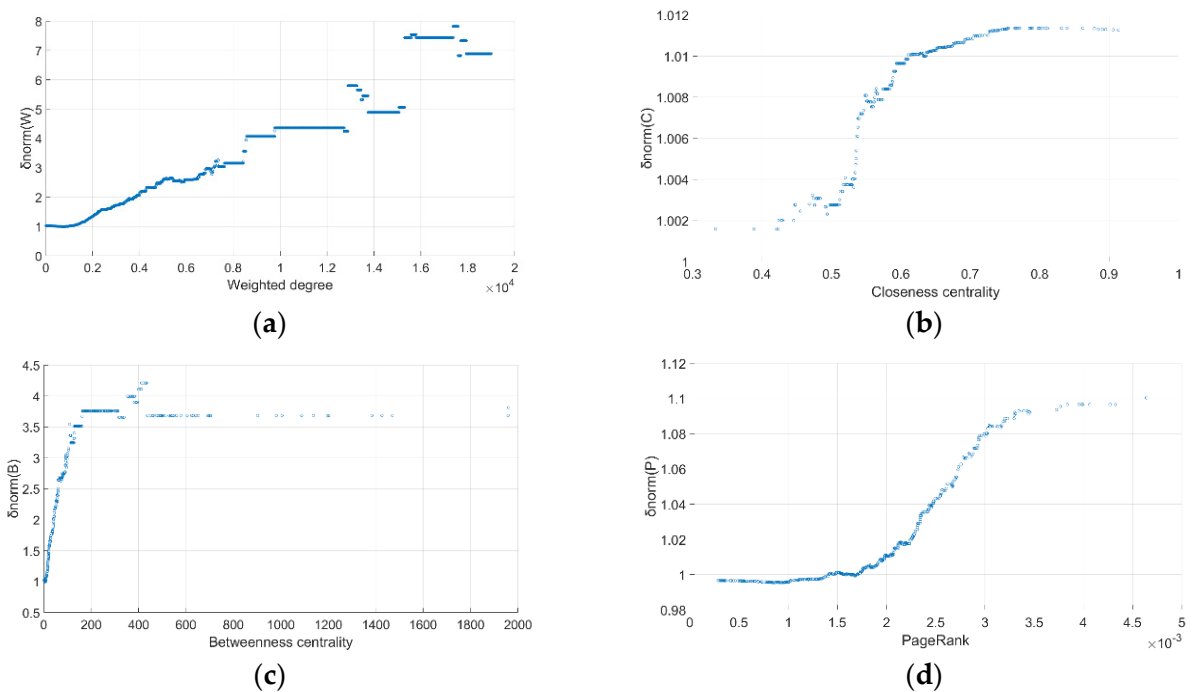
**Figure 2.** Keyword co-occurrence networks of the four topics. (a) Topic 1; (b) Topic 2; (c) Topic 3; (d) Topic 4.



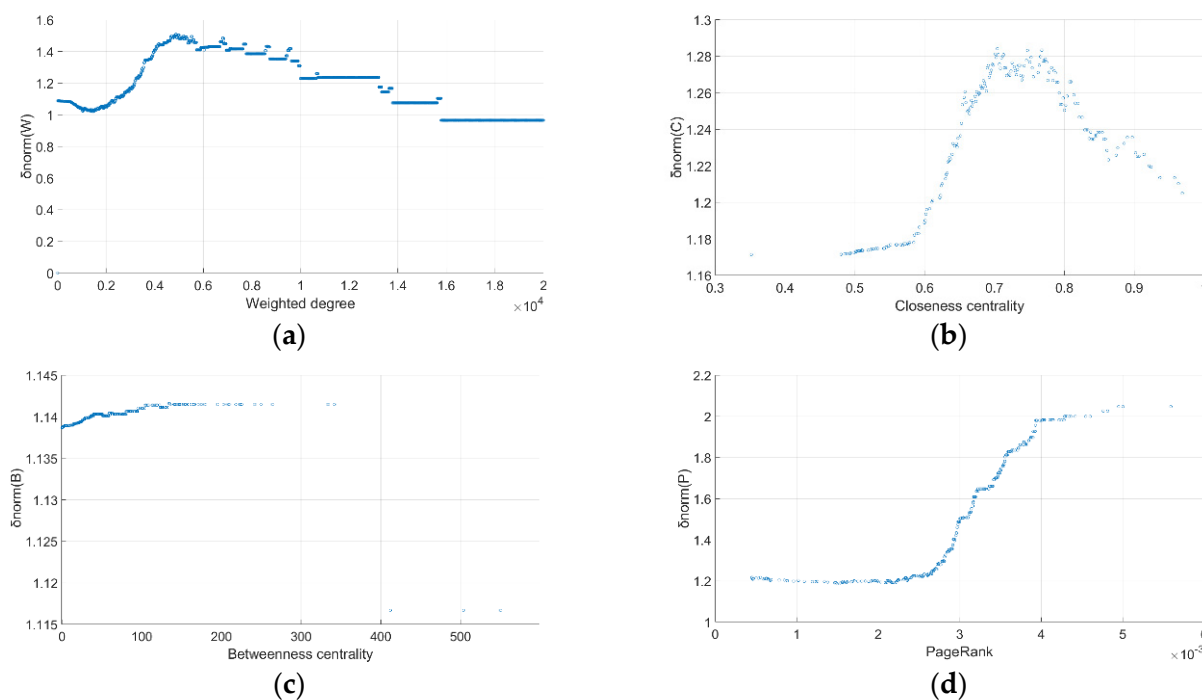
**Figure 3.** Normalized rich-club connectivity of the keyword co-occurrence networks of Topic 1 ( $\xi$  = weighted degree, closeness centrality, betweenness centrality and PageRank values). (a)  $\delta_{norm}(W)$ ; (b)  $\delta_{norm}(C)$ ; (c)  $\delta_{norm}(B)$ ; (d)  $\delta_{norm}(P)$ .



**Figure 4.** Normalized rich-club connectivity of the keyword co-occurrence networks of Topic 2 ( $\xi$  = weighted degree, closeness centrality, betweenness centrality and PageRank values). (a)  $\delta_{norm}(W)$ ; (b)  $\delta_{norm}(C)$ ; (c)  $\delta_{norm}(B)$ ; (d)  $\delta_{norm}(P)$ .



**Figure 5.** Normalized rich-club connectivity of the keyword co-occurrence networks of Topic 3 ( $\xi$  = weighted degree, closeness centrality, betweenness centrality and PageRank values). (a)  $\delta_{norm}(W)$ ; (b)  $\delta_{norm}(C)$ ; (c)  $\delta_{norm}(B)$ ; (d)  $\delta_{norm}(P)$ .



**Figure 6.** Normalized rich-club connectivity of the keyword co-occurrence networks of Topic 4 ( $\xi$  = weighted degree, closeness centrality, betweenness centrality and PageRank values). (a)  $\delta_{norm}(W)$ ; (b)  $\delta_{norm}(C)$ ; (c)  $\delta_{norm}(B)$ ; (d)  $\delta_{norm}(P)$ .

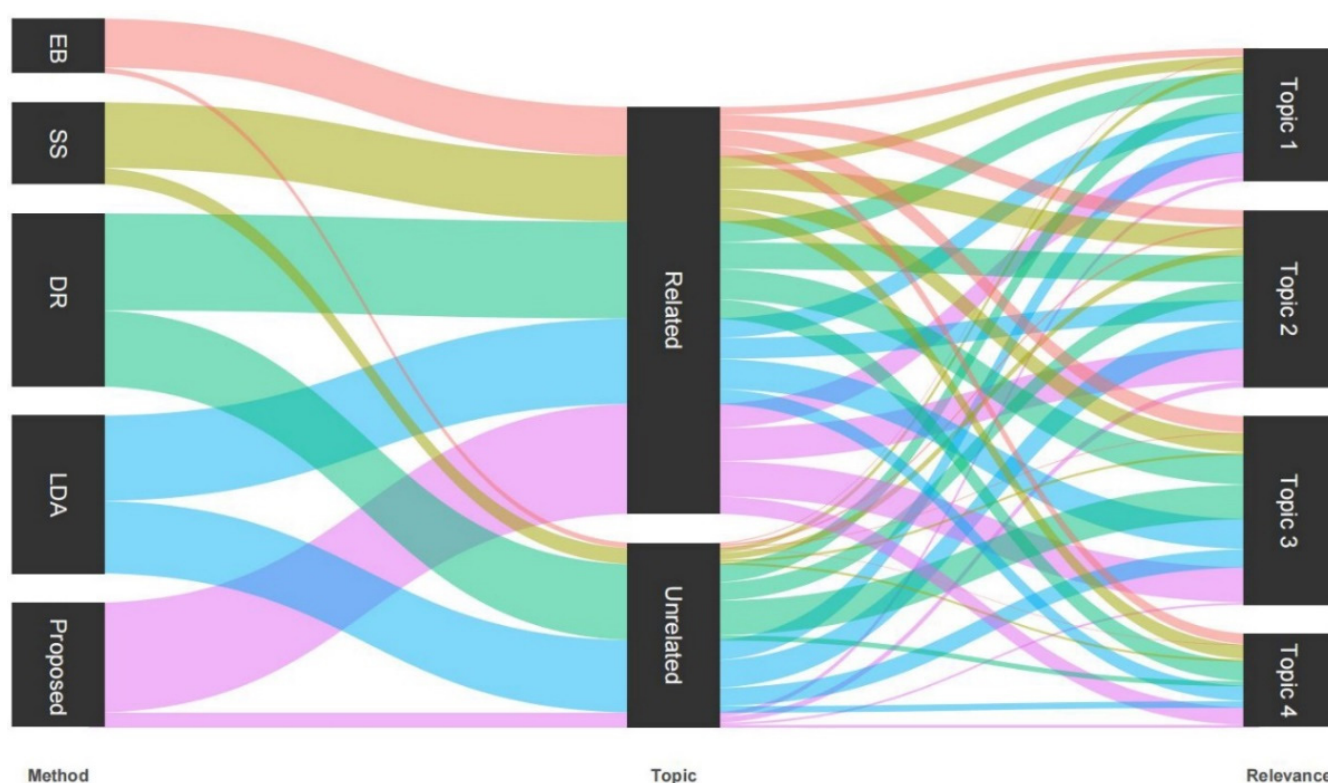
**Table 5.** Results for the four topics based on the proposed strategy.

	Additional Keywords	Number of Online Posts in Total
Topic 1	Yi Nao + punishment, blacklist + 177, rights + patients + protection, hospital + fair, physicians + safety	13,997
Topic 2	cyber violence + refuse, Human flesh + search, keyboard man + real-name system, network order + regulate	18,732
Topic 3	two sessions + 2020, Hong Kong + independence, disease + control, proposal + NPC, proposal + CPPCC, poverty + alleviation, vaccine + HPV	17,997
Topic 4	Ministry of Justice + 27, international students + right, national treatment + fair	9316

Then, 10 people who had performed social media-based studies in the past manually identified the related online posts for each topic from the collected data based on each keyword selection approach. Next, we calculated the values of the relevance coefficient  $r$  and the number of topic-related posts  $N_c$ . The final results are shown in Table 6 and Figure 7. The results show that although the size of data can be increased by using the DR and LDA methods as compared to EB and SS methods, the quality of data fell dramatically as the selected additional keywords cannot accurately represent the researched topic. In comparison, our proposed strategy could effectively address this shortcoming. In particular, although the relevance coefficient values of Topic 1, Topic 2 and Topic 4 of our proposed strategy were smaller than those of the EB method, we obtained the highest numbers of topic-related posts in three researched topics, which are 11,864, 15,873 and 16,951. In addition, the relevance coefficient for Topic 3 using our proposed strategy was even greater than that of the EB method while the size of the data was increased as much as 2.16 times. These promotions were mainly due to the use of the additional keywords that were identified using graph-based techniques.

**Table 6.** Comparison the results between the five models.

	Topic 1			Topic 2			Topic 3			Topic 4		
	$N_T$	$r$ (%)	$N_c$	$N_T$	$r$ (%)	$N_c$	$N_T$	$r$ (%)	$N_c$	$N_T$	$r$ (%)	$N_c$
<b>EB</b>	4201	86.53	3635	8226	85.57	7039	8319	93.92	7813	5404	93.67	5062
<b>SS</b>	7771	71.01	5518	13,592	76.24	10,363	10,136	88.52	8972	7930	86.22	6837
<b>DR</b>	18,904	53.27	10,070	21,527	59.66	12,843	31,074	46.54	14,462	11,926	79.07	9430
<b>LDA</b>	19,207	49.24	9458	23,248	43.09	10,018	23,256	62.34	14,498	10,359	69.61	7211
<b>Proposed strategy</b>	13,997	84.76	11,864	18,732	84.74	15,873	17,997	94.19	16,951	9316	89.81	8367

**Figure 7.** Sankey diagram for comparing the results from the EB, SS, DR, LDA and the proposed strategy.

## 6. Conclusions

Like some previous studies [27,28,43], our work highlights the important role played by the keyword selection process in social media-based studies. We posit that keyword selection should be viewed as a primary step in social media-based studies as it directly determines the quantity and quality of the collected data and thus greatly affects the reliability of the corresponding findings and the validity of any decisions made based on these findings. In previous studies that have used social media-based data, the EB, SS, DR and LDA methods were employed for keyword selection, with the former two being more commonly used [22,23,25]. A major drawback of these approaches is that they ignore the relationship between keywords, which results in the collection of a large quantity of unrelated data and the omission of related data.

Aiming at solving this problem, we proposed a new strategy for keyword selection in social media-based studies that can increase the size and ensure the high quality of collected data while improving the reliability and validity of the corresponding conclusions. This strategy considers the co-occurrence of relationships between different keywords for a particular researched topic. We used the normalized rich-club connectivity considering

the weighted degree, closeness centrality, betweenness centrality and PageRank values to extract “rich keywords”, and community detection was performed to identify several keyword combinations that can entirely and accurately represent a topic. In order to test the performance of our proposed strategy, four topics, namely, medical violence, cyber violence, the National People’s Congress (NPC) and Chinese People’s Political Consultative Conference (CPPCC) Annual Sessions 2020 and the New Permanent Residence Law for Foreigners (Draft Version), were considered in an empirical experiment. The results have shown that if quality is assured, our proposed strategy can greatly improve the quantity of the collected data, which is considerably important for social governance [68]. In addition, while the use of graph-based techniques can be helpful for keyword selection, we still highlight the primary and indispensable importance of the role of scholar experience. In other words, unlike previous studies simply emphasizing the design of quantitative methods [43,45] in keyword selection tasks, we suggest that scholars should use their personal experience to optimize and adjust their keyword results based on qualitative methods, such as expert diagnosis and consultant systems.

In sum, the evidence proves that our proposed strategy outperforms all baselines in all the four topics, which demonstrates the high effectiveness of our proposed strategy in keyword selection for social media-based analysis studies. Moreover, it should be noted that our empirical experiment only considered issues regarding original online posts. However, comment data also provide a main data source in social media-based studies [69]. In general, as the comment data are opinions directed toward a particular original online post, if the relevance of the original online post can be ensured, the relevance of its comment data can be guaranteed, as well. Therefore, if the number of collected original online posts increases, the number of comment data can further increase in time and finally improve the size and quality of the data for analyzing.

The contributions of our main findings in addressing challenges of sustainability are listed below. First, as social media has been one of the main data sources for exploring public perceptions and behaviors of issues about sustainability [70], such as environmental pollution [71], sustainable education [72] and city development [73], our proposed keyword selection method could enhance the quality of the collected data from social media and accordingly yield a better and more accurate understanding of the public opinions and behaviors. As a result, more effective solutions for addressing environmental and social sustainability problems could be designed. Second, social media has been commonly viewed as an important channel for spreading knowledge about sustainability science and concepts [74]. As our proposed model can offer useful guidelines for select the most important and relevant keywords from social media data, it contributes to knowledge discovery and influencing factors identification of knowledge dissemination. Third, as the keyword is one of the primary bases for fake news detection [75,76], which is an important part of sustainable education and emergency management, our proposed model can increase the detection accuracy by improving keyword selection process. Fourth, keyword selection is important in bibliometric analytics [77]. Therefore, our proposed model is helpful for providing a more accurate and comprehensive map considering existing literatures related to sustainability. Overall, the contribution of this article is of great significance and has a wide range of applications for future studies focusing on sustainability.

There are, however, some limitations to this study. One major limitation of the current study is that we focused only on Weibo. We believe that similar shortcomings in keyword selection for social media-based data analysis can be found on Twitter and Facebook. Therefore, in our future studies, we will perform more empirical experiments focusing on Twitter and Facebook in consideration of the characteristics of their generated online posts. Moreover, since a purely quantitative approach for keyword selection may be insufficient, we will focus on how to combine the experience of scholars with quantitative methods more effectively and efficiently in the future in order to discover new strategies that can more precisely identify keywords for analysis.



**Author Contributions:** Conceptualization, S.H. and Y.L.; methodology, Y.L. and X.L.; software, X.D.; validation, S.H., Y.L. and X.L.; formal analysis, Y.L.; investigation, Y.L. and X.L.; resources, X.D.; data curation, X.L.; writing—original draft preparation, Y.L.; writing—review and editing, S.H.; visualization, X.L.; supervision, S.H.; project administration, S.H.; funding acquisition, S.H. and X.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (71904010) and the Fundamental Research Funds for the Central Universities (CUC210C002).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data used to support the findings of this study are available from the corresponding author upon request.

**Acknowledgments:** The authors thank the reviewers for their useful discussions and comments on this manuscript.

**Conflicts of Interest:** No conflict of interest exists in the submission of this manuscript, and the manuscript is approved by all authors for publication. We have no relevant financial interests in this manuscript.

## References

1. Duari, S.; Bhatnagar, V. Complex Network based Supervised Keyword Extractor. *Expert Syst. Appl.* **2020**, *140*, 112876. [\[CrossRef\]](#)
2. Noh, H.; Jo, Y.; Lee, S. Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Syst. Appl.* **2015**, *42*, 4348–4360. [\[CrossRef\]](#)
3. Lu, W.; Liu, Z.; Huang, Y.; Bu, Y.; Li, X.; Cheng, Q. How do authors select keywords? A preliminary study of author keyword selection behavior. *J. Inf.* **2020**, *14*, 101066. [\[CrossRef\]](#)
4. Behrouzi, S.; Sarmoor, Z.S.; Hajsadeghi, K.; Kavousi, K. Predicting scientific research trends based on link prediction in keyword networks. *J. Inf.* **2020**, *14*, 101079. [\[CrossRef\]](#)
5. Cheng, Q.; Wang, J.; Lu, W.; Huang, Y.; Bu, Y. Keyword-citation-keyword network: A new perspective of discipline knowledge structure analysis. *Scientometrics* **2020**, *124*, 1923–1943. [\[CrossRef\]](#)
6. Kim, A.J.; Jang, S.; Shin, H.S. How should retail advertisers manage multiple keywords in paid search advertising? *J. Bus. Res.* **2019**, *130*, 539–551. [\[CrossRef\]](#)
7. Ayanso, A.; Karimi, A. The moderating effects of keyword competition on the determinants of ad position in sponsored search advertising. *Decis. Support Syst.* **2015**, *70*, 42–59. [\[CrossRef\]](#)
8. Huang, M.-H.; Whang, T.; Xuchuan, L. The Internet, Social Capital, and Civic Engagement in Asia. *Soc. Indic. Res.* **2016**, *132*, 559–578. [\[CrossRef\]](#)
9. Hong, S.; Nadler, D. Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Gov. Inf. Q.* **2012**, *29*, 455–461. [\[CrossRef\]](#)
10. Lee, S.; Xenos, M. Social distraction? Social media use and political knowledge in two U.S. Presidential elections. *Comput. Hum. Behav.* **2019**, *90*, 18–25. [\[CrossRef\]](#)
11. 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. [\[CrossRef\]](#)
12. Cebollero-Salinas, A.; Cano-Escoriaza, J.; Orejudo, S. Social Networks, Emotions, and Education: Design and Validation of e-COM, a Scale of Socio-Emotional Interaction Competencies among Adolescents. *Sustainability* **2022**, *14*, 2566. [\[CrossRef\]](#)
13. Mollema, L.; Harmsen, I.A.; Broekhuizen, E.; Clijnk, R.; De Melker, H.; Paulussen, T.; Kok, G.; Ruiters, R.; Das, E. Disease Detection or Public Opinion Reflection? Content Analysis of Tweets, Other Social Media, and Online Newspapers During the Measles Outbreak in the Netherlands in 2013. *J. Med. Internet Res.* **2015**, *17*, e128. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Anik, M.A.H.; Sadeek, S.N.; Hossain, M.; Kabir, S. A framework for involving the young generation in transportation planning using social media and crowd sourcing. *Transp. Policy* **2020**, *97*, 1–18. [\[CrossRef\]](#)
15. Wu, D.; Cui, Y. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decis. Support Syst.* **2018**, *111*, 48–59. [\[CrossRef\]](#)
16. Samaddar, S.; Roy, S.; Akter, F.; Tatano, H. Diffusion of Disaster-Preparedness Information by Hearing from Early Adopters to Late Adopters in Coastal Bangladesh. *Sustainability* **2022**, *14*, 3897. [\[CrossRef\]](#)
17. Dong, T.; Liang, C.; He, X. Social media and internet public events. *Telemat. Inform.* **2017**, *34*, 726–739. [\[CrossRef\]](#)
18. Salleh, S.M. From Survey to Social Media: Public Opinion and Politics in the Age of Big Data. *J. Comput. Theor. Nanosci.* **2017**, *23*, 10696–10700. [\[CrossRef\]](#)
19. Hoffmann, M.; Heft, A. “Here, There and Everywhere”: Classifying Location Information in Social Media Data—Possibilities and Limitations. *Commun. Methods Meas.* **2020**, *14*, 184–203. [\[CrossRef\]](#)

20. Abkenar, S.B.; Kashani, M.H.; Mahdipour, E.; Jameii, S.M. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telemat. Inform.* **2021**, *57*, 101517. [[CrossRef](#)]
21. Luo, Q.; Zhai, X. "I will never go to Hong Kong again!" How the secondary crisis communication of "Occupy Central" on Weibo shifted to a tourism boycott. *Tour. Manag.* **2017**, *62*, 159–172. [[CrossRef](#)] [[PubMed](#)]
22. Liu, X.; Hu, W. Attention and sentiment of Chinese public toward green buildings based on Sina Weibo. *Sustain. Cities Soc.* **2018**, *44*, 550–558. [[CrossRef](#)]
23. Su, L.; Stepchenkova, S.; Kirilenko, A.P. Online public response to a service failure incident: Implications for crisis communications. *Tour. Manag.* **2019**, *73*, 1–12. [[CrossRef](#)]
24. D'Andrea, E.; Ducange, P.; Bechini, A.; Renda, A.; Marcelloni, F. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Syst. Appl.* **2019**, *116*, 209–226. [[CrossRef](#)]
25. Han, X.; Wang, J. Using Social Media to Mine and Analyze Public Sentiment during a Disaster: A Case Study of the 2018 Shouguang City Flood in China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 185. [[CrossRef](#)]
26. Cody, E.M.; Reagan, A.J.; Mitchell, L.; Dodds, P.; Danforth, C.M. Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll. *PLoS ONE* **2015**, *10*, e0136092. [[CrossRef](#)]
27. Wang, S.; Chen, Z.; Liu, B.; Emery, S. Identifying search keywords for finding relevant social media posts. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 3052–3058.
28. Zheng, X.; Sun, A. Collecting event-related tweets from twitter stream. *J. Assoc. Inf. Technol.* **2019**, *70*, 176–186. [[CrossRef](#)]
29. Lian, Y.; Liu, Y.; Dong, X. Strategies for controlling false online information during natural disasters: The case of Typhoon Mangkhut in China. *Technol. Soc.* **2020**, *62*, 101265. [[CrossRef](#)]
30. Di Tommaso, G.; Gatti, M.; Iannotta, M.; Mehra, A.; Stilo, G.; Velardi, P. Gender, rank, and social networks on an enterprise social media platform. *Soc. Netw.* **2020**, *62*, 58–67. [[CrossRef](#)]
31. Lian, Y.; Dong, X.; Liu, Y. Topological evolution of the internet public opinion. *Phys. A Stat. Mech. Its Appl.* **2017**, *486*, 567–578. [[CrossRef](#)]
32. Xu, K.; Qi, G.; Huang, J.; Wu, T.; Fu, X. Detecting bursts in sentiment-aware topics from social media. *Knowl. Based Syst.* **2018**, *141*, 44–54. [[CrossRef](#)]
33. Liu, W.; Lai, C.-H.; Xu, W. Tweeting about emergency: A semantic network analysis of government organizations' social media messaging during Hurricane Harvey. *Public Relat. Rev.* **2018**, *44*, 807–819. [[CrossRef](#)]
34. Alam, M.; Abid, F.; Guangpei, C.; Yunrong, L. Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications. *Comput. Commun.* **2020**, *154*, 129–137. [[CrossRef](#)]
35. Qian, X.; Li, M.; Ren, Y.; Jiang, S. Social media based event summarization by user–text–image co-clustering. *Knowl. Based Syst.* **2019**, *164*, 107–121. [[CrossRef](#)]
36. Ibrahim, N.F.; Wang, X. Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Comput. Hum. Behav.* **2019**, *96*, 32–45. [[CrossRef](#)]
37. Kang, G.; Ewing-Nelson, S.R.; Mackey, L.; Schlitt, J.T.; Marathe, A.; Abbas, K.; Swarup, S. Semantic network analysis of vaccine sentiment in online social media. *Vaccine* **2017**, *35*, 3621–3638. [[CrossRef](#)] [[PubMed](#)]
38. Wu, F.; Huang, Y.; Song, Y.; Liu, S. Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decis. Support Syst.* **2016**, *87*, 39–49. [[CrossRef](#)]
39. Asif, M.; Ishtiaq, A.; Ahmad, H.; Aljuaid, H.; Shah, J. Sentiment analysis of extremism in social media from textual information. *Telemat. Inform.* **2020**, *48*, 101345. [[CrossRef](#)]
40. Wang, Y.; Li, H.; Wu, Z. Attitude of the Chinese public toward off-site construction: A text mining study. *J. Clean. Prod.* **2019**, *238*, 117926. [[CrossRef](#)]
41. Cai, B.; Geng, Y.; Yang, W.; Yan, P.; Chen, Q.; Li, D.; Cao, L. How scholars and the public perceive a "low carbon city" in China. *J. Clean. Prod.* **2017**, *149*, 502–510. [[CrossRef](#)]
42. Zhang, M.; Liu, X.; Xia, Y. Online Public Opinion Alienation Analysis of Significant Doctor-patient Dispute Cases: Taking Xiangtan Pregnant Woman Event as an Example. *J. Intell.* **2016**, *35*, 64–69.
43. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X. Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [[CrossRef](#)]
44. He, J.; Li, L.; Wang, Y.; Wu, X. Targeted aspects oriented topic modeling for short texts. *Appl. Intell.* **2020**, *50*, 2384–2399. [[CrossRef](#)]
45. Yin, X.; Wang, H.; Yin, P.; Zhu, H.; Zhang, Z. A co-occurrence based approach of automatic keyword expansion using mass diffusion. *Scientometrics* **2020**, *124*, 1885–1905. [[CrossRef](#)]
46. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [[CrossRef](#)]
47. Newman, M.E.J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2004**, *69*, 066133. [[CrossRef](#)]
48. Wagenseller, P.; Wang, F.; Wu, W. Size Matters: A Comparative Analysis of Community Detection Algorithms. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 951–960. [[CrossRef](#)]
49. Zhou, S.; Mondragon, R. The Rich-Club Phenomenon in the Internet Topology. *IEEE Commun. Lett.* **2004**, *8*, 180–182. [[CrossRef](#)]
50. Tang, C.; Dong, X.; Lian, Y.; Tang, D. Do Chinese hospital services constitute an oligopoly? Evidence of the rich-club phenomenon in a patient referral network. *Futur. Gener. Comput. Syst.* **2019**, *105*, 492–501. [[CrossRef](#)]

51. Ball, G.; Aljabar, P.; Zebari, S.; Tusor, N.; Arichi, T.; Merchant, N.; Robinson, E.C.; Ogundipe, E.; Rueckert, D.; Edwards, A.D.; et al. Rich-club organization of the newborn human brain. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7456–7461. [[CrossRef](#)]
52. Wei, Y.; Song, W.; Xiu, C.; Zhao, Z. The rich-club phenomenon of China’s population flow network during the country’s spring festival. *Appl. Geogr.* **2018**, *96*, 77–85. [[CrossRef](#)]
53. Smilkov, D.; Kocarev, L. Rich-club and page-club coefficients for directed graphs. *Phys. A Stat. Mech. Appl.* **2010**, *389*, 2290–2299. [[CrossRef](#)]
54. Ren, T.; Wang, Y.-F.; Du, D.; Liu, M.-M.; Siddiqi, A. The guitar chord-generating algorithm based on complex network. *Phys. A Stat. Mech. Appl.* **2016**, *443*, 1–13. [[CrossRef](#)]
55. Kim, D.-J.; Min, B.-K. Rich-club in the brain’s macrostructure: Insights from graph theoretical analysis. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1761–1773. [[CrossRef](#)] [[PubMed](#)]
56. Cinelli, M. Generalized rich-club ordering in networks. *J. Complex Netw.* **2019**, *7*, 702–719. [[CrossRef](#)]
57. Lv, L.; Zhang, K.; Zhang, T.; Bardou, D.; Zhang, J.; Cai, Y. PageRank centrality for temporal networks. *Phys. Lett. A* **2019**, *383*, 1215–1222. [[CrossRef](#)]
58. Salavati, C.; Abdollahpouri, A.; Manbari, Z. Ranking nodes in complex networks based on local structure and improving closeness centrality. *Neurocomputing* **2019**, *336*, 36–45. [[CrossRef](#)]
59. Jin, H.; Zhang, C.; Ma, M.; Gong, Q.; Yu, L.; Guo, X.; Gao, L.; Wang, B. Inferring essential proteins from centrality in interconnected multilayer networks. *Phys. A Stat. Mech. Appl.* **2020**, *557*, 124853. [[CrossRef](#)]
60. Maslov, S.; Sneppen, K. Specificity and Stability in Topology of Protein Networks. *Science* **2002**, *296*, 910–913. [[CrossRef](#)]
61. Azevedo, F.A.C.; Carvalho, L.R.B.; Grinberg, L.T.; Farfel, J.M.; Ferretti, R.E.L.; Leite, R.E.P.; Filho, W.J.; Lent, R.; Herculano-Houzel, S. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **2009**, *513*, 532–541. [[CrossRef](#)]
62. Herculano-Houzel, S. The human brain in numbers: A linearly scaled-up primate brain. *Front. Hum. Neurosci.* **2009**, *3*, 31. [[CrossRef](#)]
63. Opsahl, T. Structure and Evolution of Weighted Networks. Ph.D. Thesis, Queen Mary, University of London, London, UK, 2009.
64. Traag, V.A. Faster unfolding of communities: Speeding up the Louvain algorithm. *Phys. Rev. E* **2015**, *92*, 032801. [[CrossRef](#)]
65. Wang, G.; Liu, Y.; Li, J.; Tang, X.; Wang, H. Superedge coupling algorithm and its application in coupling mechanism analysis of online public opinion supernetwork. *Expert Syst. Appl.* **2015**, *42*, 2808–2823. [[CrossRef](#)]
66. Kaplan, S.; Vakili, K. The double-edged sword of recombination in breakthrough innovation: The Double-Edged Sword of Recombination. *Strateg. Manag. J.* **2015**, *36*, 1435–1457. [[CrossRef](#)]
67. Bastani, K.; Namavari, H.; Shaffer, J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Syst. Appl.* **2019**, *127*, 256–271. [[CrossRef](#)]
68. Zakharchenko, A.; Peráček, T.; Fedushko, S.; Syerov, Y.; Trach, O. When Fact-Checking and ‘BBC Standards’ Are Helpless: ‘Fake Newsworthy Event’ Manipulation and the Reaction of the ‘High-Quality Media’ on It. *Sustainability* **2021**, *13*, 573. [[CrossRef](#)]
69. Lyu, Y.; Chow, J.C.-C.; Hwang, J.-J. Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China’s social media Weibo. *Child. Youth Serv. Rev.* **2020**, *116*, 105250. [[CrossRef](#)]
70. Mangla, M.; Ambarkar, S.; Akhare, R.; University of Mumbai; Computer Department at LTCOE. A study to Analyze impact of social media on society: WhatsApp in particular. *Int. J. Educ. Manag. Eng.* **2020**, *10*, 1–10. [[CrossRef](#)]
71. Shutaleva, A.; Martyushev, N.; Nikonova, Z.; Savchenko, I.; Abramova, S.; Lubimova, V.; Novgorodtseva, A. Environmental Behavior of Youth and Sustainable Development. *Sustainability* **2021**, *14*, 250. [[CrossRef](#)]
72. Sobaih, A.E.E.; Hasanein, A.; Elshaer, I.A. Higher Education in and after COVID-19: The Impact of Using Social Network Applications for E-Learning on Students’ Academic Performance. *Sustainability* **2022**, *14*, 5195. [[CrossRef](#)]
73. Castro, A.I.G.; López, L.J.R. Sustainability and Resilience of Emerging Cities in Times of COVID-19. *Sustainability* **2021**, *13*, 9480. [[CrossRef](#)]
74. Borah, P.S.; Iqbal, S.; Akhtar, S. Linking social media usage and SME’s sustainable performance: The role of digital leadership and innovation capabilities. *Technol. Soc.* **2022**, *68*, 101900. [[CrossRef](#)]
75. Mostafa, G.; Ahmed, I.; Junayed, M.S. Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data. *Int. J. Inf. Technol. Comput. Sci.* **2021**, *13*, 38–48. [[CrossRef](#)]
76. Akinyemi, B.; Adewusi, O.; Oyebade, A. An Improved Classification Model for Fake News Detection in Social Media. *Int. J. Inf. Technol. Comput. Sci.* **2020**, *12*, 34–43. [[CrossRef](#)]
77. Bielański, M.; Korbiel, K.; Taczanowska, K.; Pardo-Ibañez, A.; González, L.-M. How tourism research integrates environmental issues? A keyword network analysis. *J. Outdoor Recreat. Tour.* **2022**, *37*, 100503. [[CrossRef](#)]