Technology in Society 71 (2022) 102112

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Public attitudes on open source communities in China: A text mining analysis

Shengjie Hou, Xiang Zhang, Biyi Yi, Yi Tang *

National Innovation Institute of Defense Technology, Beijing, 100071, PR China

ARTICLE INFO

Keywords:
Open source community
Public engagement
Social media
Text mining
Social network analysis

ABSTRACT

The Development of open source community has been a major concern in many countries. Sustainable public engagement is one of the essential conditions for establishing an open source community. Social media offers a new channel for understanding public opinions on OSCs. This paper conducts a content analysis focusing on social media data regarding the OSCs in China. Topic clustering, sentiment classification, and social network analysis are used to analyze the text data. Results show that most people on social media support the development of OSCs in China, but there are still some objections that believe open source will reduce the innovation ability of China. Based on the findings, we suggest that the government should fully understand public attitudes on OSCs and respond in time to the public by using social media. More high-quality China's independent OSCs should be built to enable Chinese local open source contributors to survive and communicate with each other. In addition, we suggest that a comprehensive and reasonable evaluation and incentive system and more effective copyright protection measures should be created. Overall, this paper contributes to the field of development of OSCs from the perspective of users.

1. Introduction

According to a significant body of research on open and distributed innovation, innovation processes are often open and diffused, across organizational and geographic barriers [1–4], which well matches the core function of the Open Source Communities (OSCs). OSCs can be defined as a kind of community for s knowledge-sharing and production [5]. As early as 1980s, Stallman [6] argued that computer programs should be considered a public good, and he has called for the development of free software and the establishment of the free software Foundation. Since then, the concept of free and open source community has gained increasing traction among developers and users alike, and the proliferation of OSCs has had a significant impact on the open source platform ecosystem as a whole. The open nature of OSCs enables users to tap into their own creativity and collaborate to build low-cost, high--performance, and high-customer-satisfaction items more quickly than conventional design techniques [7]. In specific, through the OSCs, users were able to customize and improve the program to meet their specific requirements, and subsequently approved their programs into the community source code base.

Literature focusing on OSCs mainly regarded issues about its applications [3,8,9] and improvement [2,10]. For the first part, OSCs have gained tremendous popularity as effective external sources in various areas, including but not limited to corporate governance [11], technology spreading [12], project management [13], and public administration [14]. For the second part, a few studies proposed some improvement measures from the theoretical perspective [2,10,15,16]. For instance, Shaikh and Vaast [2] built a process theory of how transparency and openness are balanced with opacity and closure in OSCs according to the needs of development work. In addition, there is a consensus existing in related studies, which is that in order to obtain continuous and actively input and support, OSCs developers must establish strong connections with their users in order for projects to be sustainable [10,17–20]. This is extremely important for the survival of OSCs. Moreover, some studies found that sustaining the OSCs is the most challenge [18] as OSS developers rarely raise issues and questions about the OSCs and generally ignored how to benefit the users of the community. This mainly because the shortage in the access of user engagement in improving the community [20,21]. Thus, a deep understanding of users' comments on OSCs refers to essential evidence for OSCs improvement. However, there have been few studies in this regard in past. For instance, Dennehy et al. categorized 20,651 discursive

* Corresponding author.
E-mail address: xtangee@hotmail.com (Y. Tang).

https://doi.org/10.1016/j.techsoc.2022.102112
Received 4 January 2022; Received in revised form 31 August 2022; Accepted 3 September 2022
Available online 13 September 2022
0160-791X/© 2022 Elsevier Ltd. All rights reserved.

instances of inconsistencies that happened in a big developing OSC's patch review comments, and they also accordingly proposed some recommendations on sustaining OSCs [10]. Ferreira et al. used data from OSC mailing list to investigate how community sentiment impacts the development process of open source software [22].

Many online social platforms, including as Twitter, Facebook, and Weibo, have emerged as a result of the Internet's expansion, allowing members of the public to freely express their own ideas [23,24]. In general, when confronted with particular circumstances, internet users' own emotions, attitudes, and behavioral patterns are referred to as online public opinions [25]. By examining continuity, sample sizes, and relevance to study issues, Jiang et al. [26] outlined the key distinctions between IPOs and regular survey data. They also looked at IPO data from three perspectives: post intensity, sentiment analysis, and subject identification. Their findings gave a more systematic knowledge of online public opinion analysis. Additionally, sentiment and viewpoints are two primary social media-based analysis.

In terms of OSCs in China, most existing studies are empirical analysis regarding the present situation and application aspects. For example, Wu et al. [27] conducted an excellent work on describing the development status of Chinese open source software industry and ecosystem. They found that the technology selection would be limited because of the strong support of Chinese government. The findings of this paper contributed to the development of open source in Chinese enterprises and individual hobbyists. Huang et al. [28] provided novel evidence focusing on the application of open innovation Chinese firms with different sizes by using the Tobit regression approach, in which a comparison analysis was conducted. Chen et al. [29] explored the knowledge sharing mechanism open source software projects in China based on 403 valid surveys. They found that social network is a primary factor in China that should be improved by strengthening the collaboration between software enterprises and OSCs. In addition, some studies focused on use of open source data in solving real world problems [30–33]. For instance, Kpozehouen et al. [30] researched the application of open source data to identify early signals of pneumonia in China prior to official recognition of the COVID-19 outbreak. Zhou and Xu [31] used open source data to detect the walking routes environment of urban parks in China. In sum, compared with western countries, the development of OSCs in China is still at a low stage. Therefore, we believe that a better knowledge for improving Chinese OSCs from the perspective of the public is required.

The main purpose of this paper is to provide a content analysis focusing on online public opinions regarding the OSCs in China. Some text mining approaches, such as text clustering algorithm and sentiment classification algorithm, were applied to analyze the collected data. We intended to figure out the main problems regarding OSCs in China and accordingly propose some useful suggestions.

The rest of the present paper is structed as follows: In Section 2, we present the research method, including data collection, and several text mining approaches. Section 3 shows the results. Section 4 concludes the paper and put forwards some practical suggestions on the development of OSC in China.

2. Research method

2.1. Data

"OSCs" was the keywords utilized to gather the data. A web crawler technology called Octopus was utilized to gather all online postings from November 1, 2020 to November 1 of 2021 on Weibo. Weibo is one of China's most popular social media networks. According to the 45th China Statistical Report on Internet Development (CNNIC) issued by the China Internet Network Information Center in March 2020, more than half of all Chinese netizens use Weibo on a monthly basis. We collected data every day of the research period in order to get as many relevant postings as possible. Original, reposted, and commented posts were included in the data set. For each post, we recorded the user ID, username, user description, number of followings, number of followers, number of posts, location, user labels, URL, posting time, text of the post, number of reposts, number of comments, and number of likes for each user. Then, the SPSS 25.0 data cleaning tool was used to eliminate incomplete samples from the raw data during data reprocessing. Advertisements and other posts that were not related to the OSCs were manually deleted depending on their content. Finally, a total of 78 popular original posts were collected, together with the associated commenting and reposting data of which have the number of commenting data and reposting data higher than 1000. Additionally, regular expression operations in R were used to remove noise from the acquired data, which increased the textual analysis' performance. The total number of online posts for analyzing is 4901. It can be noticed that the number of our collected data is relatively small, which indicate that the attention of the public on the development of OSCs in China is under a low degree. Thus, since the active participation of users is an important prerequisite for the sustainable development of OSCs, the government should pay more attention on strengthening the publicity of OSCs.

According to Dong et al. [25], a constructed analysis framework for analyzing social media data has been proposed, which contained three steps: trend and spatial analysis, topic clustering, and sentiment classification. In addition to these three steps, we also focused on the network characteristics of reposting posts by employing social network analysis approach. Thus, our applied framework contained four steps, shown as Fig. 1.

2.2. Social network analysis

Social network analysis is a method to analyze the data from the perspective of the network, which has been applied in various fields, including but not limited to medical [34], information spreading [35], ecological [36], and transportation [37]. For existing studies focusing on the information spreading on social media, reposting network was frequently considered, in order to identify the influential users that dominated the discussion related to the topic [38–40]. For instance, Jain and Katarya proposed a Louvain-based hybrid network analysis method to identify the top-N local and global opinion leader within the community and social network [40]. Chen et al. put forward an integrated bounded confidence model to investigate the interaction mechanism between opinion leaders and their followers, in which the social network analysis approach was used to build the opinion dynamics environment [41]. In the present paper, a particular reposting network is represented as $\boldsymbol{G}=(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}=\{v_1, v_2, \ldots, v_N\}$ is the set of nodes, and $\boldsymbol{E}=\{e_1, e_2, \ldots, e_M\}$ represents the edges of the network $\boldsymbol{G}$. In the network, a node represents a user and an edge represents a reposting relationship. Then, a weighted PageRank method proposed by Zhang [42] was used to determine the importance ranking of users contained in the reposting network.

2.3. Text mining approaches

In the present paper, three text mining approaches, namely, word segmentation, topic clustering, and sentiment analysis were applied.

2.3.1. Word segmentation

For texts in Chines, unlike those in English, there is a need to carry out the word segmentation before analyzing the data by using text mining approaches. *Jieba* is a R program that has demonstrated to be successful in processing the contents of Chinese social media data in a number of past experiments [43–45]. Thus, *Jieba* was used in the present paper. In addition, we utilized a vocabulary given by Sougou Pinyin for word segmentation. Sougou Pinyin is a Chinese character input technique that is extensively used across the globe. It enables users of various skill levels to learn or practice Chinese typing. Its many real-time updatable lexicons have largely contributed to the method's success.

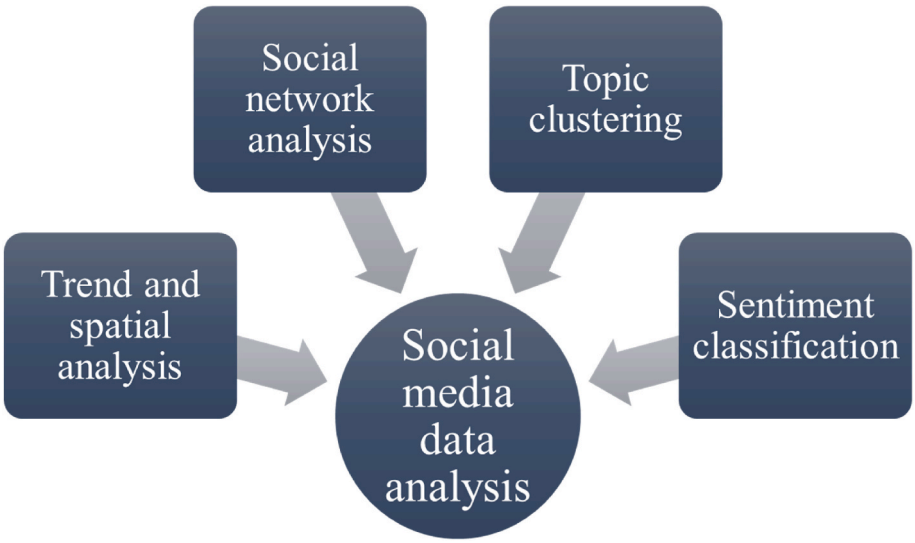

Fig. 1. Research framework.

Then, the stop words that lack actual meaning or contribute little to an argument, such as "in", "on", and "that" [46], were cleaned up by employing the stop word list provided by *Jieba*.

2.3.2. Topic clustering

Existing research have presented a variety of topic clustering techniques, with Latent Dirichlet Allocation (LDA) being one of the most often used. This article employed LDA to execute text analysis for topic clustering about the acquired data because it has shown effective performance in topic clustering analysis in many previous studies [47–49]. The statistical association of words given in investigated papers without regard for word order is the foundation of LDA [49]. The k-dimensional topic smoothing parameter and the k-dimensional word smoothing parameter are set to 0.1 and 0.01, respectively, for parameter settings. In addition, the ideal number of subjects was determined using the trial and error technique. Although this technique would take a long time to repeat the clustering operations with varied numbers of topics, the accuracy could be assured to a significant degree [50,51].

Furthermore, the TF-IDF approach is used to calculate the weight of each distinct word in each message, as given in formula (1).

$$W_{ij} = TF_{ij} \times ID_{ij} = TF_{ij} \times \log\left(N / DF_j\right) \quad (1)$$

where W_{ij}, TF_{ij} and ID_{ij} are the weight, word frequency, and inverse document frequency of characteristic word j in post i. N is the total number of posts, and DF_j is the number of posts that include characteristic word j, respectively.

2.3.3. Sentiment analysis

To determine how the general public felt about the researched topic, sentiment analysis was used. In general, there are two kinds of methods for sentiment classification, namely, lexicon-based approach and the supervised Machine Learning (SML) approach. The first is based on a sentiment lexicon that contains a large number of keywords along with the sentiment score, and the second requires a large amount of labeled data and a SML classifier. As the SML approaches are generally more accurate than lexicon-based approaches [52], especially for Chinese texts [53], the SML approaches were applied in the present paper. In specific, the Bi-directional Long Short-Term Memory (Bi-LSTM) is a sort of recurrent neural network utilized in sentiment analysis, and it was used in this research. The Bi-LSTM consists of a forward and backward LSTM, and it has shown excellent performance in processing Chinese Twitter data for sentiment analysis [54,55]. For word embedding, Google's Word2Vec, a word vector training tool, was employed. The hidden size of each LSTM unit was set to 300 for the Bi-LSTM, and the learning rate was set to 0.01 for optimization. To train the model, we utilized a dataset of 120,000 online postings from Weibo, half of which were tagged as positive and half of which were labeled as negative. This information may be found at https://qcsdn.com/q/a/49489.html. The generated classifier was then utilized to determine the sentiment trend in our collected online postings.

3. Results

3.1. Time and spatial analysis results

Figs. 2 and 3 depict the trend and geographical distribution of collected original postings, respectively. As can be seen in Fig. 2, there is one significant peak, which corresponds to an online post describing OpenEuler, an open source project in the field of operating system implemented by HUAWEI. The post stated that it has gathered more than 5000 developers around the world in only 18 months, attracted tens of millions of clicks in more than 1000 cities, and began to be widely implemented in finance, communication and other industries. At the regional level, Fig. 3 demonstrates that Beijing, Guangdong, Jiangsu, Anhui, and Shandong, all of which are in the top echelon of economic growth and Internet development, have the most related online posts.

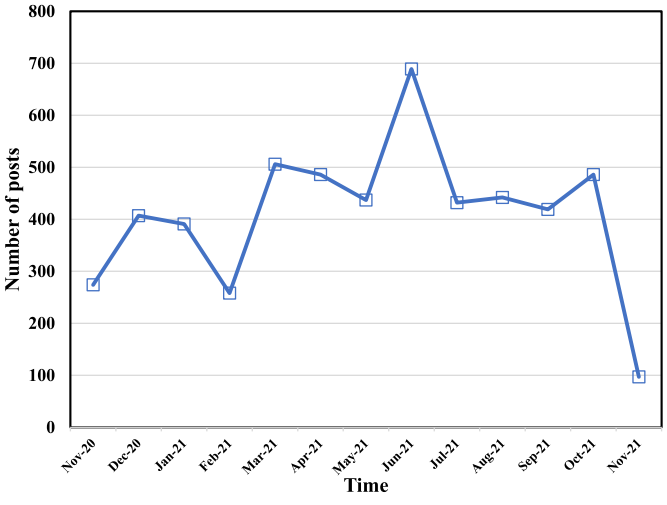

Fig. 2. Trend of collected online posts.

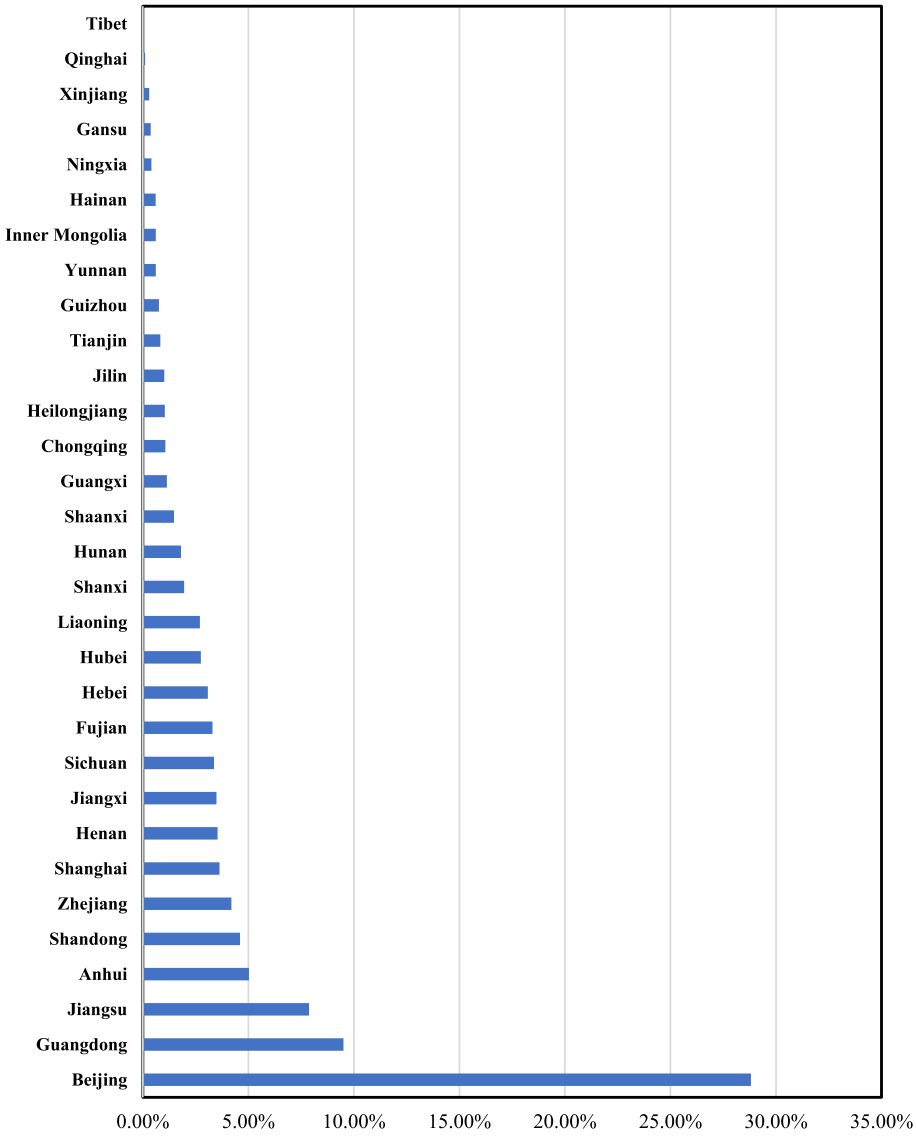

Fig. 3. Geographical distribution of collected online posts.

3.2. Reposting network analysis results

After that, we sketched up the reposting network of online posts from Weibo. This network is a weighted directed network, in which the weight is measured by the number of reposting relations between nodes in the network. This network is seen in Fig. 4. As seen in Fig. 5, a logarithmic plot depicts the weighted degree distribution of the reposting network that was developed. Its distribution function was calculated as $P(k) = 0.8219{*}k^{-2.941}$, with statistical significance at less than one percent chance of error; further, the R-square and adjusted R-square values were, respectively, 0.9998 and 0.9998, indicating a power-law structure. In order to have a better understanding of the significance of each node included inside the reposting network, we also performed a weighted PageRank analysis using the approach given by Zhang [42]. For the purposes of this work, we set the parameter for managing the performance of the algorithm to be 0.85. Table 1 shows the top 40 nodes with the highest PageRank values, ranked from highest to lowest. According to Table 1, 15 of the top 20 nodes with the highest PageRank values have been recognized as being official media users. However, this proportion decreases to 50% and 45% with respect to the case of top 30 and top 40, respectively. This finding indicates that official media users have played an important role in leading the dissemination of information on OSCs, while some unofficial users also made great contributions, to a large extent.

3.3. Text mining analysis results

The word segmentation method was carried out for each examined post using *jieba* package and the results are given in Fig. 6. The size of the keyword was set based on its TF-IDF value: a bigger keyword has a larger TF-IDF value. It can be noticed that despite some keywords that are directly related to our research topic, such as Open Source and China, have a greater frequency, those with higher frequencies mostly contain Huawei, code, Protect, Software, System, and so on. This conclusion reflects the public's concern about the OSCs in China. In specific, Huawei Corporation is a main force for driving the development of OSCs in China. For instance, openEuler is an open and free OSC platform based on the Linus, which was developed by HUAWEI. In addition, open source system and open source software are two main topics discussed by the public on social media.

Then, we analyzed the collected commenting data using text mining approaches, and the results are shown in Fig. 7. For commenting data, we also extracted top 100 keywords based on the ranking results of the TF-IDF value, which show similar results regarding the original collected

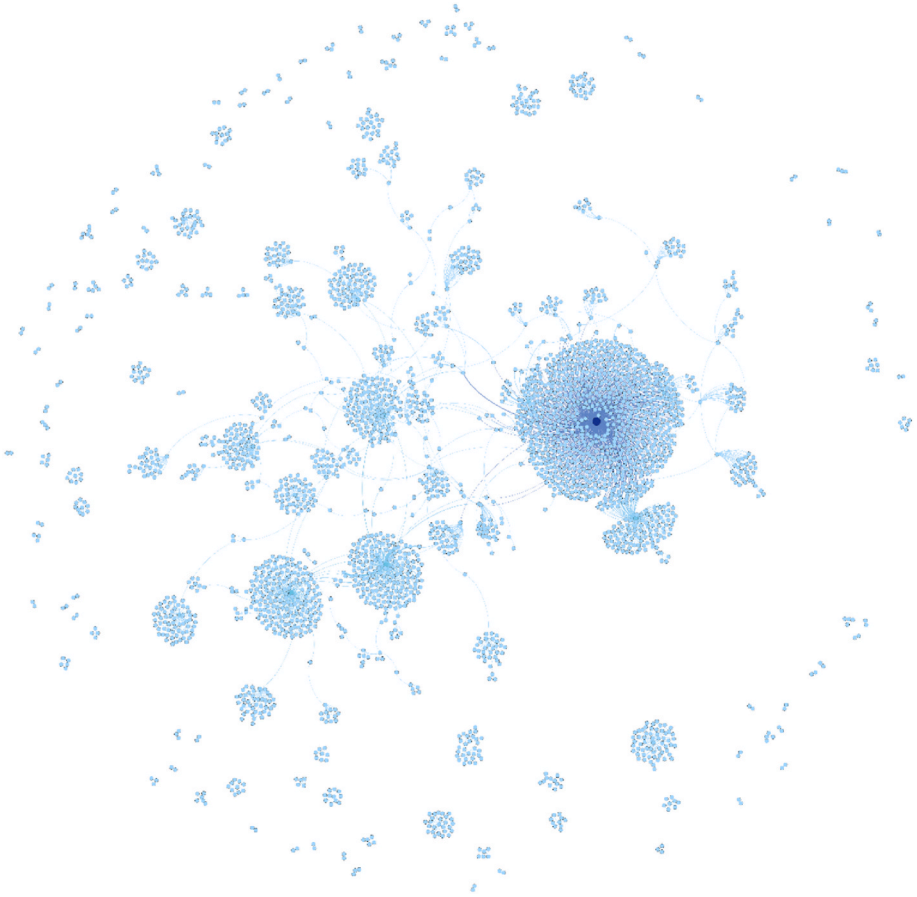

Fig. 4. Reposting network of collected online posts (a directed weighted network).

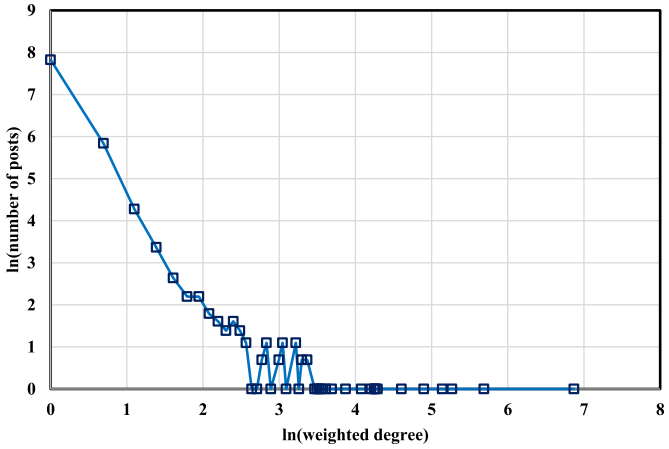

Fig. 5. Degree distribution of weighted degree.

data. For sentiment analysis results, the proportion of negative comments and positive comments is 33.36% and 66.64%, respectively. In order to further analyze public views focused on OSCs in China, three topics with negative tendency and three topics with positive tendency were extracted from the commenting data using the LDA approach and the trial and error technique, shown as follows.

3.3.1. Topics with positive tendency

The first positive topic is "Salute developers who committed to building Chinese OSCs". Peoples holding this topic describe personal respects and appreciation for the strength, dedication, and innovations

Table 1
Top 40 nodes with the largest weighted PageRank value.

Order	User Name	PageRank	Order	User Name	PageRank
1	驻**馆	0.001985	21	O**胸	0.000666
2	徐**意	0.001985	22	大**水	0.000662
3	f**3	0.001704	23	奇**素	0.000655
4	P**区	0.001199	24	a**d	0.00065
5	a**h	0.001088	25	以**德	0.000641
6	陈**D	0.000931	26	M**E	0.00064
7	豆**球	0.000844	27	浦**米	0.00064
8	烟**d	0.000814	28	C**Sui	0.000639
9	包**岗	0.000808	29	文**忙	0.000633
10	天**凶	0.000789	30	大**车	0.000626
11	c**o	0.000786	31	方**狼	0.000626
12	丸**姨	0.000772	32	M**V	0.000626
13	孟**S	0.000771	33	應**將	0.000608
14	西**篱	0.00077	34	金**多	0.000604
15	默**闻	0.000708	35	苏**店	0.000598
16	一**灵	0.000706	36	s**7	0.000597
17	m**n	0.000683	37	章**I	0.000592
18	w**l	0.00067	38	軒**慶	0.00059
19	K**着	0.000667	39	叶**米	0.000587
20	蓝**e	0.000666	40	m**6	0.000586

Note: To protect privacy, "*" is used to hide user names.

of individuals and organizations on establishing and promoting Chinese OSCs. For example, some comments stated:

> "I think these developers have excellent technology and they could promote the development of technology. Thank you!"

> "These developers are really powerful and have paid a lot of sweat behind them!"

Fig. 6. Word segmentation results of collected original online posts. Note: size of the keyword is positively related to its TF-IDF value.

The second positive topic contains posts related to supporting OSCs projects. Most of peoples who holding this topic are coders, as they generally viewed themselves as the members of programmer communities. They thought that China should speed up the construction of its own open source community, because it is important to improve original innovation ability. Some people stated that:

> "The project of open source community is really good. It belongs to the programmer community!"

> "We firmly support the establishment of local open source communities"

The majority of postings in the third positive topic are about supporting the operating system developed by China. Compared to the second positive topic, this one pays more attention on the operating system. This is because that openEuler is one of the most typical and conspicuous open source products in China at present. As an example, one comment stated:

> "I think open source projects in the field of operating system will promote the development of China's economy. I'm looking forward to it."

> "As a development log in the operating system, openeuler will have more than 5000 developers around the world in only 18 months. This is really powerful."

3.3.2. *Topics with negative tendency*

The first negative topic is "Condemn the practice of applying for a patent with open source code". According to their perspectives, using open source codes to apply patents is an illegal act, and people who did this should be removed from the OSCs and punished by the law. In addition, we discovered that the majority of postings containing information regarding this issue support the establishment of OSCs in China. Some typical comments of this topic are as follows:

> "It's shameless to apply for a patent with open source software!"

> "Resist patent trolls and maintain the open source environment"

The second negative topic is "Worry that open source community will lead enterprises not to innovate". People holding this topic do not support promoting the establishment of the OSCs in China because they believe that although the OSCs provide a lot of valuable codes and products and convenient interfaces for people, OSCs will make people over rely on existing open source codes. Consequently, people will not create new coders, but to modify or apply directly old ones, which will greatly reduce the ability of original innovation of China. Some typical comments of this topic are as follows:

> "Open source software will make people lazy and don't innovate. Enterprises without underlying architecture can't survive!"

> "That's interesting. Doesn't open source mean you don't have to innovate?"

The third negative topic is "Worry about the development status of China's local open source community". This topic is similar with our identified second positive topic, because they both support OSCs projects in China. The difference is that the related posts of the former express negative sentiment by showing worried mental state. They argue that China should build its own OSCs as soon as possible in order to promote the development of software industry:

> "China does not have its own open source community to maintain and manage these codes. In case of our irresistible factors, all our software industries in China will be very dangerous."

4. Discussion and practical suggestions

OSCs can break the discipline and professional barriers in the field of various areas, including but not limited to technological innovation, software development, military science, e-government, and electronic commerce [56]. It has also been viewed as one of the most important indicators of original innovation ability [57,58]. Thus, how to enhance the development of OSCs has been a big concern of the government across the globe. Many previous studies focused on how to increase the number of users participating in these communities, but few carried out the research from the perspective of users themselves. We claim that sustainable users' participation is the premise of building OSCs. Timely and comprehensively understand the core need of users and potential users of OSCs is of great significance. In this regard, it seems that social media is an efficient means of attaining one's goals.

In this paper, data related to OSCs from social media in China were collected and analyzed. In specific, time and spatial analysis was carried out to show the trend and geographical distribution of the collected data by visualization. The results show that there is one significant peak in the timeline and most related messages were posted by people in the regions with better performance in economic growth and Internet development, such as Beijing and Guangdong. According to previous studies [59,60], temporal and spatial analysis is of great significance in understanding the evolution of the researched events, such as natural disasters, which could provide valuable basis to support decision-makings of the government. With respect to OSCs, we believe that the research regarding the relationship between attention level of the public and OSCs development degree could be a worthwhile discussion for future studies. Then, social network analysis approach was

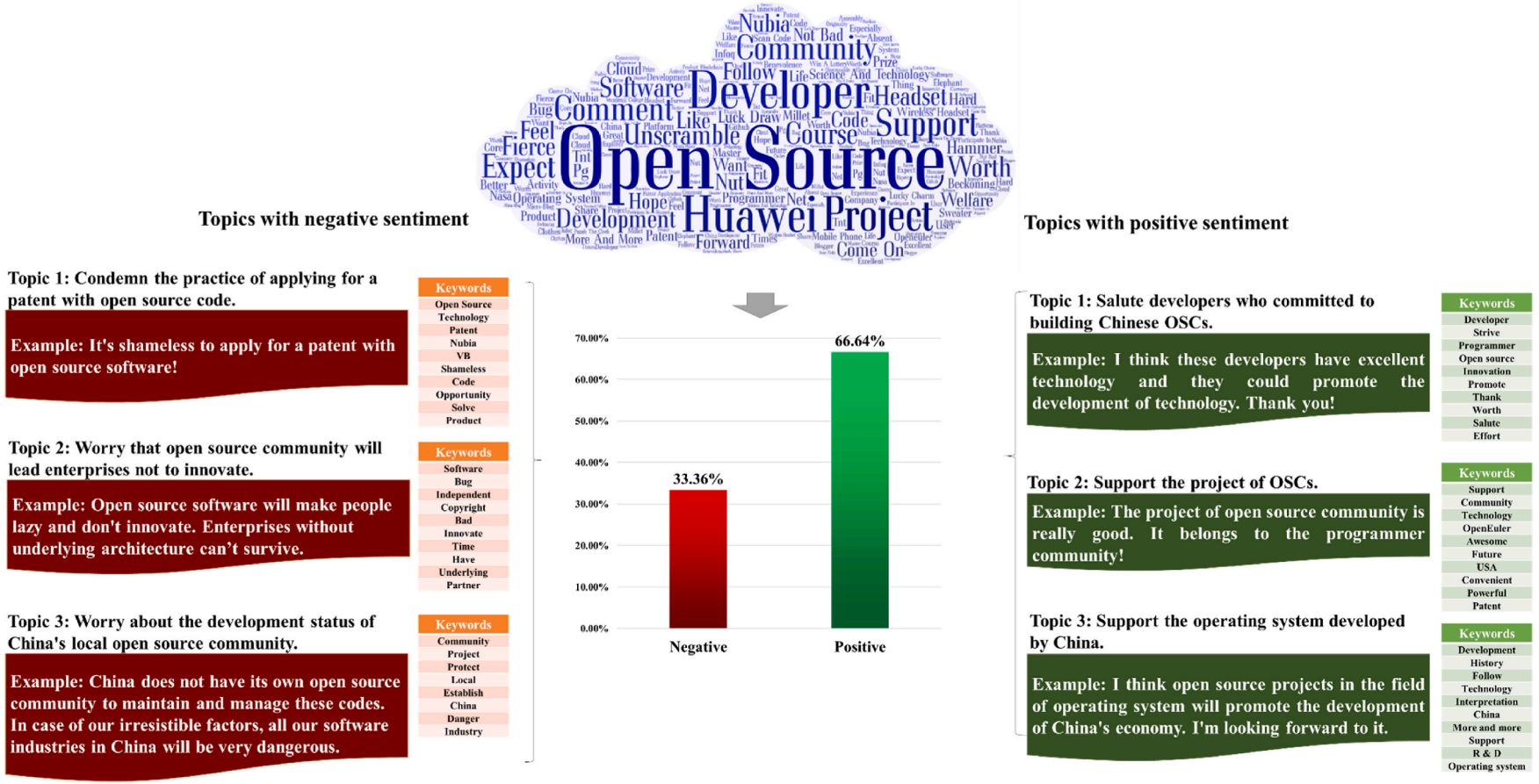

Fig. 7. Text mining results of collected commenting online posts.

used to identify important users in spreading related information. We also combined sentiment classification method and topic modelling technique to extract public's primary views on relevant issues with different sentiment tendency. We discovered that the majority of the public in social media apps support to promote the establishment of Chinese own OSCs, and some individuals even believe that it should be implemented as soon as possible. However, some people held a different viewpoint because they thought that OSCs will make people heavily rely on the open source resources, and accordingly arouse adverse impacts on the original innovation ability of China.

Compared to previous studies that applied traditional survey method to investigate public attitudes toward issues related to OSCs [7,61], the current study could provide a more comprehensive, complete, and real picture, to a large extent, as the use of social media data and data mining techniques enabled us to obtain much more samples' opinions. Moreover, different from the existing studies that only considered developers or users of OSCs [29,62], we expanded the sample to nearly all people who posted opinions related to OSCs on social media, and they may become potential developers or potential users of OSCs in the future, which is an important part of driving forces for sustainable development of OSCs. In addition, this paper revealed some new evidence on clarifying how do the people response to the Chinese development of OSCs and what is the most dissatisfied about OSCs of the public in China. We believe that such findings may be used for the government to guide programs aimed at creating pleasant interactions among members of the OSCs, establishing social connections, and assisting the community in completing its goals.

To better improve the development of OSCs in China, this research offers following recommendations based on our findings. These recommendations are also useful for other countries or regions that focus on OSCs.

i. As a first step, the government and organizations that are responsible for building and promoting OSCs could utilize social media to better understand the public opinions about OSCs. Social media can also be an effective tool for propaganda and educate by offering fostering two-way communication and the sharing of information with the general public [63]. The government should devote greater attention to the causes of public opinions with negative tendency. According to our findings, these negative opinions mainly concentrated in the adverse effects of OSCs on the innovation ability. Thus, the government could increase the interpretation and publicity of the OSCs in improving innovation ability by using social media and some other means in online and offline. Based on our findings obtained from reposting network analysis, official media in social media refers to the main force spreading information related to OSCs. In addition to official media, the government should increase the publicity enthusiasm of enterprises, organizations or even related individuals by introducing some incentives. Moreover, teenagers, as the largest potential contributor of digital technologies in the future [64], should be regarded as one of the main objects of publicity and education.
ii. More high-quality China's independent OSCs like GitHub should be built by the government or technology-related large companies, such as HUAWEI and Open Source China. According to the Octo-verse report of GitHub in 2019, China has been the second contributor of OSCs.[1] This data raised a question that why were Chinese local contributors more willing to participate in the enhancement of GitHub than in Chinese OSCs. This is partly because the rare of Chinese self-developed OSCs. OSCHINA is a Chinese OSC with a relatively long history, which was founded in August2, 008.[2] It has more than 3 million registered members in the world. OpenEuler is a typical example representing the current development state of China's OSCs. The aim of openEuler is to create an innovation platform, build a unified and open operating system supporting multi-processing architecture, and promote the prosperity and development of software and hardware application ecology through community cooperation. However, it mainly focuses on the operating system. In addition, in 2020, the Ministry of industry and information technology of China announced that Gitee.com[3] was selected as the foundation platform for building China's independent OSC.[4] Gitee.com is a git-based code hosting service platform launched by the Open Source China. It also provides an open source software release and communication community for developers to carry out technical exchanges and communications. However, compared with GitHub, the world's largest OSC, there is still much space for OSCHINA and Gitee.com to progress. Thus, the government should increase investment and policy support for OSCs development. In this aspect, the government could take the lead in setting up more open source software foundations by encouraging domestic giant enterprises such as Huawei, Baidu, Alibaba and Tencent to establish professional foundations around key and excellent domestic open source projects.
iii. As users are the main force driving the sustainable development of OSCs, a comprehensive and reasonable evaluation and incentive system should be created [65]. The system could fully mobilize the innovation enthusiasm of contributors such as universities, scientific research institutes and enterprises, to improve the allocation efficiency of innovation resources [66], and provide a strong internal driving force for the development of national science and technology. From the macro level, the government should build an open source talent training and evaluation system by encouraging universities and professional training institutions to set up theoretical and practical training courses related to open source software. In addition, a reward for open source talent could be set up, to enable excellent open source talents to get reasonable returns, and release their innovation vitality for the development of China's local OSCs. From the micro level, for the OSCs themselves, blockchain technology could be applied to the underlying architecture of OSCs, as it is helpful for realizing participatory governance. The goal of participatory governance is to ensure that the contributors' trust in the decision-making group is actually to give the decision-makers ownership, and make the responsibilities decentralized rather than centralized as much as possible.
iv. According to your findings, many people express negative emotions on the behavior that using open source code for patent application in social media. Behind this finding is the public's doubts about the ability of open source data protection. In addition, many previous studies found that code plagiarism has become one of the main threats to the OSCs [67–69]. Thus, both the government and OSCs builders should pay more attention on the copyright protection of open source resources. It should be highlighted that open source could be free to use but it does not mean "brings the principle". That is to say, the legal subject status of OSCs developers and contributors is unshakable and undeniable. From the policy perspective, the government should release more targeted and detailed copyright protection laws and regulations for open source resources [70], and improve the awareness of OSCs users to protect their achievements. From the

[1] https://cloud.tencent.com/developer/news/471279.

[2] https://www.oschina.net/home/aboutosc.

[3] https://gitee.com/.

[4] https://baijiahao.baidu.com/s?id=1675998443724732627&wfr=spider&for=pc.

technical perspective, blockchain is an effective tool for copyright protection due to its distributed storage mode and advanced encryption technology [71]. For instance, Jing et al. [72] proposed a blockchain-based code copyright management system that can provide a better level of storage security for OSCs.

5. Conclusion

This paper investigated public opinions about OSCs in China. Related social media data was used to provide a content analysis. Topic clustering, sentiment classification, and social network analysis approaches were employed to analyze the data. The results show that most people in social media support the development of Chinese own OSCs, but there are still some people disagree because they believed that OSCs will let China reduce the original innovation ability. Based on the findings, several suggestions for promoting Chinese OSCs were proposed, including: i. the government and organizations that are responsible for promoting OSCs should fully employ social media to better understand the needs and dissatisfaction of the public and accordingly put forward more useful for increasing their engagement; ii. more high-quality China's independent OSCs should be built by the government or large technology-related companies in order to provide a unified and open space for Chines local open source contributors to survive and communicate; iii. a comprehensive and reasonable evaluation and incentive system should be created in order to fully mobilize the innovation enthusiasm of contributors; iv. more attention should be paid on the copyright protection of open source resources. The paper's contribution is to provide some useful evidence that can be used as a foundation by other researchers in the field of OSCs, both in terms of how qualitative methodologies can be used to study social media data in practice and in terms of how to improve the development of OSCs from the perspective of users and potential users.

The following are the limitations of this study. First, due to Weibo's information access limitations and the low degree in public attentions on OSCs in China at present, the sample size for analyzing is relatively small. In future studies, we will collect more useful and relevant data by employing questionnaires and interviews methodologies. Second, this paper only conducted a content mining analysis to understand public opinions on OSCs in China, which did not contain any empirical analysis regarding the effects of different influencing factors on the engagement degree of contributors. Thus, we will build some data-driven regression models based on the social media data to conduct deeper insight into the development of OSCs.

Authorship statement

Hou Shengjie: Conceptualization, Methodology, Software, Investigation, Formal analysis, Software, Writing-original draft, Writing-review & editing, Funding acquisition. Xiang Zhang: Methodology, Software, Writing-original draft. Biyi Yi: Methodology, Formal analysis, Writing-review & editing. Yi Tang: Validation, Conceptualization, Resources, Writing-original draft, Visualization, Writing-review & editing, Supervision, Project administration.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgement

Thankfulness shall be expressed to the reviewers for their useful discussions and comments on this manuscript. The present paper is supported by the Beijing Philosophy and Social Science Foundation [20JCC021].

References

[1] L. Dahlander, M. Magnusson, How do firms make use of open source communities? Long. Range Plan. 41 (6) (2008) 629–649.
[2] M. Shaikh, E. Vaast, Folding and unfolding: balancing openness and transparency in open source communities, Inf. Syst. Res. 27 (4) (2016) 813–833.
[3] M.R. Guertler, N. Sick, Exploring the enabling effects of project management for SMEs in adopting open innovation – a framework for partner search and selection in open innovation projects, Int. J. Proj. Manag. 39 (2) (2021) 102–114.
[4] E.K.R.E. Huizingh, Open innovation: state of the art and future perspectives, Technovation 31 (1) (2011) 2–9.
[5] M. AlMarzouq, V. Grover, J.B. Thatcher, Taxing the development structure of open source communities: an information processing view, Decis. Support Syst. 80 (2015) 27–41.
[6] R. Stallman, Why Software Should Be Free?, 1994. Available from: http://www.gnu.ai.mit.edu/philosophy/shouldbefree.html.
[7] H. Baytiyeh, J. Pfaffman, Open source software: a community of altruists, Comput. Hum. Behav. 26 (6) (2010) 1345–1354.
[8] M. Germonprez, et al., A theory of responsive design: a field study of corporate engagement with open source communities, Inf. Syst. Res. 28 (1) (2017) 64–83.
[9] J. Yang, et al., Energy spectral behaviors of communication networks of open-source communities, PLoS One 10 (6) (2015) e0128251-e0128251.
[10] D. Dennehy, et al., Sustaining Open Source Communities by Understanding the Influence of Discursive Manifestations on Sentiment, Information systems frontiers, 2020.
[11] E. Piva, F. Rentocchini, C. Rossi-lamastra, Is open source software about innovation? Collaborations with the open source community and innovation performance of software entrepreneurial ventures, J. Small Bus. Manag. 50 (2) (2012) 340–364.
[12] A. Hemetsberger, et al., Learning and knowledge-building in open-source communities: a social-experiential approach, Manag. Learn. 37 (2) (2006) 187–214.
[13] N.J. Foss, L. Frederiksen, F. Rullani, Problem-formulation and problem-solving in self-organized communities: how modes of communication shape project behaviors in the free open-source software community, Strat. Manag. J. 37 (13) (2016) 2589–2610.
[14] J. Longo, T.M. Kelley, GitHub use in public administration in Canada: early experience with a new collaboration tool, Can. Publ. Adm. 59 (4) (2016) 598–623.
[15] J. Jagtiani, C. Bach, C. Huntley, Leveraging big data from open source to improve software project management, IEEE Eng. Manag. Rev. 46 (1) (2018) 65–79.
[16] V. Achal, C.S. Chin, Building Materials for Sustainable and Ecological Environment, Springer Singapore Pte. Limited, Singapore, 2021.
[17] M. Ozer, D. Vogel, Contextualized relationship between knowledge sharing and performance in software development, J. Manag. Inf. Syst. 32 (2) (2015) 134–161.
[18] M.M. Appleyard, H.W. Chesbrough, The dynamics of open strategy: from adoption to reversion, Long. Range Plan. 50 (3) (2017) 310–321.
[19] M. Shaikh, Negotiating open source software adoption in the UK public sector, Govern. Inf. Q. 33 (1) (2016) 115–132.
[20] J. Song, C. Kim, What is needed for the sustainable success of OSS projects: efficiency analysis of commit production process via git, Sustainability 10 (9) (2018) 3001.
[21] J. Gamalielsson, B. Lundell, Sustainability of Open Source software communities beyond a fork: how and why has the LibreOffice project evolved? J. Syst. Software 89 (1) (2014) 128–145.
[22] J. Ferreira, et al., Sentiment Analysis of Open Source Communities: an Exploratory Study, ACM, 2019.
[23] D. Lai, et al., Addressing immediate public coronavirus (COVID-19) concerns through social media: utilizing Reddit's AMA as a framework for Public Engagement with Science, PLoS One 15 (10) (2020) e0240326-e0240326.
[24] B.W. Wirtz, et al., Public social media services: a citizen's perspective, Publ. Perform. Manag. Rev. 43 (6) (2020) 1342–1358.
[25] X. Dong, Y. Lian, A review of social media-based public opinion analyses: challenges and recommendations, Technol. Soc. 67 (2021), 101724.
[26] J. Bian, et al., Mining twitter to assess the public perception of the "internet of things, PLoS One 11 (7) (2016) e0158450-e0158450.
[27] Q. Wu, K. Klincewicz, K. Miyazaki, Analysis of the open source software sector in China, Asian J. Technol. Innovat. 14 (2) (2006) 117–141.
[28] F. Huang, J. Rice, N. Martin, Does open innovation apply to China? Exploring the contingent role of external knowledge sources and internal absorptive capacity in Chinese large firms and SMEs, J. Manag. Organ. 21 (5) (2015) 594–613.
[29] X. Chen, et al., Managing knowledge sharing in distributed innovation from the perspective of developers: empirical study of open source software projects in China, Technol. Anal. Strat. Manag. 29 (1) (2017) 1–22.
[30] E.B. Kpozehouen, et al., Using open-source intelligence to detect early signals of COVID-19 in China: descriptive study, JMIR public health and surveillance 6 (3) (2020) e18939-e18939.

[31] Z. Zhou, Z. Xu, Detecting the pedestrian shed and walking route environment of urban parks with open-source data: a case study in nanjing, China, Int. J. Environ. Res. Publ. Health 17 (13) (2020) 4826.
[32] N. Kshetri, A. Schiopu, Government policy, continental collaboration and the diffusion of open source software in China, Japan, and South Korea, J. Asia Pac. Bus. 8 (1) (2007) 61–77.
[33] P. Gu, et al., Using open source data to measure street walkability and bikeability in China: a case of four cities, Transport. Res. Rec. 2672 (31) (2018) 63–75.
[34] J. Ernst, et al., Burnout, depression and anxiety among Swiss medical students – a network analysis, J. Psychiatr. Res. 143 (2021) 196–201.
[35] X. Lin, K.A. Lachlan, P.R. Spence, Exploring extreme events on social media: a comparison of user reposting/retweeting behaviors on Twitter and Weibo, Comput. Hum. Behav. 65 (2016) 576–581.
[36] B. Jiang, et al., China's ecological civilization program–Implementing ecological redline policy, Land Use Pol. 81 (2019) 111–114.
[37] B. Li, et al., Estimation of regional economic development indicator from transportation network analytics, Sci. Rep. 10 (1) (2020), 2647-2647.
[38] M. Shi, A.C. Wojnicki, Money talks … to online opinion leaders: what motivates opinion leaders to make social-network referrals? J. Advert. Res. 54 (1) (2014) 81–91.
[39] Y. Zhao, et al., Understanding influence power of opinion leaders in e-commerce networks: an opinion dynamics theory perspective, Inf. Sci. 426 (2018) 131–147.
[40] L. Jain, R. Katarya, Discover opinion leader in online social network using firefly algorithm, Expert Syst. Appl. 122 (2019) 1–15.
[41] J. Chen, et al., Influence identification of opinion leaders in social networks: an agent-based simulation on competing advertisements, Inf. Fusion 76 (2021) 227–242.
[42] F. Zhang, Evaluating journal impact based on weighted citations, Scientometrics 113 (2) (2017) 1155–1169.
[43] C. Luo, et al., Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: semantic network analysis of two social media platforms in the United States and China, Telematics Inf. 65 (2021), 101712.
[44] Y. Lian, X. Dong, Exploring social media usage in improving public perception on workplace violence against healthcare workers, Technol. Soc. 65 (2021), 101559.
[45] Y. Lian, Y. Liu, X. Dong, Strategies for controlling false online information during natural disasters: the case of Typhoon Mangkhut in China, Technol. Soc. 62 (2020), 101265.
[46] Z. Wu, et al., A topical network based analysis and visualization of global research trends on green building from 1990 to 2020, J. Clean. Prod. 320 (2021), 128818.
[47] Y. Guo, S.J. Barnes, Q. Jia, Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation, Tourism Manag. 59 (2017) 467–483.
[48] S. Tirunillai, G.J. Tellis, Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation, J. Market. Res. 51 (4) (2014) 463–479.
[49] H. Jelodar, et al., Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimed. Tool. Appl. 78 (11) (2018) 15169–15211.
[50] K. Bastani, H. Namavari, J. Shaffer, Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Expert Syst. Appl. 127 (2019) 256–271.
[51] X. Dong, Y. Lian, The moderating effects of entertainers on public engagement through government activities in social media during the COVID-19, Telematics Inf. 66 (2022), 101746-101746.
[52] N. Mukhtar, M.A. Khan, N. Chiragh, Lexicon-based approach outperforms supervised machine learning approach for Urdu sentiment analysis in multiple domains, Telematics Inf. 35 (8) (2018) 2173–2183.
[53] W. Li, et al., Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification, Neurocomputing 387 (2020) 63–77.
[54] G. Cao, et al., Analysis of social media data for public emotion on the Wuhan lockdown event during the COVID-19 pandemic, Comput. Methods Progr. Biomed. 212 (2021), 106468-106468.
[55] M. Ling, et al., Hybrid neural network for sina Weibo sentiment analysis, IEEETtransact. computat. Soc. Syst. 7 (4) (2020) 983–990.
[56] A. Sigfridsson, A. Sheehan, On qualitative methodologies and dispersed communities: reflections on the process of investigating an open source community, Inf. Software Technol. 53 (9) (2011) 981–993.
[57] T.Y. Tang, G.J. Fisher, W.J. Qualls, The effects of inbound open innovation, outbound open innovation, and team role diversity on open source software project performance, Ind. Market. Manag. 94 (2021) 216–228.
[58] J. Kimpimäki, I. Malacina, O. Lähdeaho, Open and sustainable: an emerging frontier in innovation management? Technol. Forecast. Soc. Change 174 (2022), 121229.
[59] K. Yao, S. Yang, J. Tang, Rapid assessment of seismic intensity based on Sina Weibo — a case study of the changning earthquake in Sichuan Province, China, Int. J. Disaster Risk Reduc. 58 (2021), 102217.
[60] S. Shan, J. Peng, Y. Wei, Environmental Sustainability assessment 2.0: the value of social media data for determining the emotional responses of people to river pollution—a case study of Weibo (Chinese Twitter), Soc. Econ. Plann. Sci. 75 (2021), 100868.
[61] V. Singh, L. Holt, Learning and best practices for learning in open-source software communities, Comput. Educ. 63 (2013) 98–108.
[62] H. Bai, J.M. Yang, J.T. Wang, Research on the current situation of the open source community in China from the Network Perspective, in: PROCEEDINGS OF 2013 6TH INTERNATIONAL CONFERENCE ON INFORMATION MANAGEMENT, INNOVATION MANAGEMENT AND INDUSTRIAL ENGINEERING (ICIII 2013), vol. 1, 2013, pp. 431–435.
[63] A. Adikari, et al., Value co-creation for open innovation: an evidence-based study of the data driven paradigm of social media using machine learning, Int. J. Info. Manag. Data Insights 1 (2) (2021).
[64] A. Lareki, et al., Teenagers' perception of risk behaviors regarding digital technologies, Comput. Hum. Behav. 68 (2017) 395–402.
[65] X. Sun, Q. Zhang, Building digital incentives for digital customer orientation in platform ecosystems, J. Bus. Res. 137 (2021) 555–566.
[66] S. Liu, Research on token incentive mechanism of open source project - take block chain project as an example, IOP Conf. Ser. Earth Environ. Sci. 252 (2) (2019), 22029.
[67] N. Jing, Q. Liu, V. Sugumaran, A blockchain-based code copyright management system, Inf. Process. Manag. 58 (3) (2021), 102518.
[68] C.W. Zhu, A regime of droit moral detached from software copyright?–the undeath of the 'author' in free and open source software licensing, Int. J. Law Info Technol. 22 (4) (2014) 367–392.
[69] O. Johnny, M. Miller, M. Webbink, Copyright in open source software - understanding the boundaries, Int. Free Open Source Software Law Rev. 2 (1) (2010).
[70] J. Lee, Tripartite perspective on the copyright-sharing economy in China, Comput. Law Secur. Rep. 35 (4) (2019) 434–452.
[71] W. Liang, et al., Circuit copyright blockchain: blockchain-based homomorphic encryption for IP circuit protection, IEEE Transact. Emerg. Topics Comput. 9 (3) (2021) 1410–1420.
[72] M. Papoutsoglou, et al., An analysis of open source software licensing questions in Stack Exchange sites, J. Syst. Software 183 (2022), 111113.