COVID-19 infodemic on Chinese social media: A 4P framework, selective review and research directions

Jia Luo1,2,3, Rui Xue1 and Jinglu Hu2

Abstract

During the outbreak of the COVID-19 (2019 coronavirus disease), misinformation related to the virus spread rapidly online and have led to serious difficulties in controlling the disease. The term infodemic is coined to outline the bad effect from the extensive dissemination of misinformation during the outbreak. With regards to this phenomenon, the World Health Organization emphasized the need to fight against infodemic and asked all countries not only to make efforts in slowing down the spread of the COVID-19 but also in countering the risk caused by infodemic. Due to its negative impact, this paper analyzes infodemic on Chinese social media at the initial stage of the COVID-19 outbreak and presents a 4P framework standing for the four features of Chinese infodemic: Prevention Attention, Problem Orientation, Patterns Interaction and Points Globalization. Furthermore, a selective review of existing datasets in the neural networks domain is synthesized based on the 4P framework. Finally, research directions, including recommendations, about constructing a large-scale dataset for Chinese infodemic automatic detection are proposed.

Keywords

COVID-19, infodemic data, Chinese social media, conceptual framework, literature review

Introduction

It is well known that every outbreak is accompanied by an amount of misinformation. With the development of internet, this phenomenon is amplified by social media. Since December 2019, the COVID-19 has spread around the world. At the same time, the related misinformation is spread rapidly through social media platforms. The term infodemic1 is coined to outline the bad effect from the extensive dissemination of misinformation during the outbreak. When the COVID-19 is firstly identified in China, infodemic is even more prominent due to a lack of known information at the beginning. WeChat has become a primary news source for Chinese internet users. Moreover, a mini app named “Jiaozhen” is launched by WeChat for debunking misinformation. Hot ambiguous topics at the moment are gathered by this mini app and are labeled as true, false or questionable. After observing the data for one year, it is easy to figure out in Figure 1 that the COVID-19 outbreak has generated a tsunami of hot ambiguous topics since the end of January 2020. To further investigate the composition of hot ambiguous topics which are above 100, the data are displayed inside a pie chart as seen in Figure 2. Regarding to the labels attached by “Jiaozhen,” false and questionable hot ambiguous topics account for 77% and 15% respectively, which can be considered as main components of infodemic when the COVID-19 outbreaks in China.

Many governments and social media service providers have paid great attention to collect and verify infodemic with a lot of human efforts. However, the manual annotation has to face an issue of time delay while “timing” becomes as a new challenge as social media make infodemic go faster and further.1 Therefore, it is critical to develop artificial intelligence techniques for infodemic detection and use them to debunk infodemic automatically. Some neural network approaches have been exploited to defeat misinformation on Chinese social media. Ma et al.2 presented a

1. College of Economics and Management, Beijing University of Technology, Beijing, China
2. Graduate School of Information, Production, and Systems, Waseda University, Kitakyushu, Japan
3. International Research Fellow of Japan Society for the Promotion of Science

Corresponding author:
Rui Xue, College of Economics and Management, Beijing University of Technology, No. 100 Pingleyuan, Chaoyang District, Beijing, 100124, China.
Email: xue.rui.bjut@hotmail.com
recurrent neural networks-based method for learning continuous representations of microblog events to identify rumors. Yu et al.\textsuperscript{3} proposed a convolutional approach for misinformation identification, which can flexibly extract key features. Jin et al.\textsuperscript{4} studied a novel recurrent neural network with an attention mechanism to fuse multimodal features for effective rumor detection. Wang et al.\textsuperscript{5} displayed an event adversarial neural network to derive event-invariant features and thus benefited the detection of fake news on newly arrived events. Yu et al.\textsuperscript{6} developed an attention-based convolutional approach for misinformation identification from massive and noisy microblog posts.

Datasets are an integral part of neural network approaches. Most datasets utilized for Chinese rumor automatic detection are built using Sina Weibo, the primary microblog service in China. Some general methods have been taken for data collection and annotation. Ma et al.\textsuperscript{2} collected 2313 known rumor data from Sina community management center and 2351 non-rumor events by crawling the posts of general threads that were not reported as rumors. Wu et al.\textsuperscript{7} used the same way to gather 2601 false rumors and 2536 normal messages where each had at least 100 reposts. Jin et al.\textsuperscript{4} gathered data from authoritative sources where rumor posts came from the official rumor debunking system of Sina Weibo and non-rumor tweets came from Xinhua News Agency, an authoritative news agency in China. Unlike previous datasets that were focused on only text content, original texts were collected with attached images and available surrounding social contexts. Moreover, duplicated images were removed by a near-duplicated image detection algorithm based on locality sensitive hashing while very small or long images were also eliminated.

A few preliminary works\textsuperscript{8} attempted to train neural networks on rumors and non-rumors data from Sina Weibo for the sake of easy access and great number of labeled data. However, the precision rate was as low as 28% when they were used for infodemic detection. It is easy to tell that the poor performance comes from the mismatching between general rumors and infodemic. Due to the shortage of large expert annotation datasets, neural network approaches do not play an important role for detecting infodemic during the COVID-19 outbreak. As datasets are an integral part of neural network approaches, this paper aims to analyse Chinese infodemic during the COVID-19 outbreak to present a framework, selective review and recommendations for constructing Chinese infodemic datasets. The main contributions of our work are summarized as follows:

1. To the best of our knowledge, the first framework of Chinese infodemic is presented, which consists of four features: Prevention Attention, Problem Orientation, Patterns Interaction and Points Globalization.
2. A selective review of existing datasets in neural networks domain is synthesized based on the suggested framework, focusing on their drawbacks when being utilized for infodemic automatic detection.

3. Research directions, including recommendations, about constructing a large-scale dataset for Chinese infodemic automatic detection are proposed.

The remaining sections of this paper are organized as follows. Section 2 presents the 4P framework of Chinese infodemic. Section 3 displays a selective review of existing datasets according to this framework. Section 4 provides research directions for prospective studies. Finally, Section 5 states conclusions.

4P Framework

The 4P framework stands for the four features of widespread infodemic on Chinese social media at the initial stage of the COVID-19 outbreak which are Prevention Attention, Problem Orientation, Patterns Interaction and Points Globalization. All data come from hot ambiguous topics published in the WeChat mini app “Jiaozhen” as WeChat is a primary news source for Chinese internet users and “Jiaozhen” has obtained rich experience for debunking misinformation. Since the peak in the number of Chinese hot ambiguous topics appears along with the initial stage of the COVID-19 outbreak in China (see Figure 1), this research focuses on misinformation around this period. To be precisely, 261 pieces of hot ambiguous topics from 21st January to 10th February 2020 are finally selected and any of them is labeled as either false or questionable in the WeChat mini app “Jiaozhen”.

Prevention attention

The COVID-19 outbreak is declared as a Public Health Emergency of International Concern since 30th January 2020. Because of its relevance with medical science, Chinese infodemic samples in this period are firstly divided into two types depending on the content: medical strongly related topics and medical weakly related topics. Furthermore, the medical strongly related topics are subdivided as prevention measures, general virus knowledge and treatment information subject to prevention, diagnosis and treatment in medical science. After careful analysis of the details of medical weakly related topics, this part is split into five sub-types as local measures, national measures, patient information, hospital status and others according to involved action doers. The numerical proportions of the above-mentioned types and subtypes are displayed in Figure 3 where medical strongly related topics account for 60%. Infodemic in this group regularly contains some infrequent used medical terms which are hard for the public to understand. Moreover, 43% of all topics concern prevention measures, which account for 71% of medical strongly related topics. This great proportion comes from the fear of the unknown at the initial stage of the COVID-19 outbreak. Infodemic about inappropriate prevention measures may mislead the public and increase the difficulty to control the disease.

Problem orientation

The WeChat mini app “Jiaozhen” attaches two human written keywords to each infodemic. According to the word cloud (Figure 4) generated by these keywords, it is obvious to tell that “COVID-19” has received tremendous public attention during its outbreak. Moreover, the other 224 keywords all have strictly or roughly a connection to this word. To make a further analysis, Figure 5 presents a histogram displaying the distribution of the 10 most frequently used keywords where only three of them are referred more than 10 times. The total citation number of the first ranking word, “COVID-19”, is much larger than the second one, “pneumonia”, which has confirmed the conclusion obtained from Figure 4. A great number of single-issue topics concentrated misinformation may worsen information overload and further cause psychological problems to the public. As a report stated by the Administration Center of China E-government Network, 51% of Chinese citizens felt slightly anxious during this period while 25% fairly anxious and 10% greatly anxious.

Patterns interaction

Digital information is mainly represented as texts, pictures or videos. Therefore, infodemic is classified into
three groups in this sub-section by the digital format and the statistical results are illustrated in Figure 6. Due to the relatively low cost for editing text information, it is the main form of infodemic during the COVID-19 outbreak which accounts for 73%. In contrast, only 6% of infodemic is spread in the form of videos. Although these videos are originally from other contexts while being glued with fake COVID-19 related descriptions, one video may be worth more than a thousand words. As a best compromise between the editing cost and the convictive power, infodemic in the form of images comprises 21% of the total data. Although the proportion is much less than the text form, image is the most contagious infodemic pattern during the COVID-19 outbreak. This conclusion is confirmed by a recent research which has modeled the spread of information with epidemic models. In their work, each social media platform is characterized by a basic reproduction number ($R_0$) where the value of $R_0$ for Twitter, Instagram and YouTube is 1.84, 2.25 and 1.61 respectively. Therefore, it is important for social media platforms to pay attention to infodemic in the form of images. Upon further analysis, these images can be divided into two sub-groups: text screenshot and text-included image. Being different from the traditional way that an image transmit information iconographically, the embedded texts play more important roles while graphic elements only enhance the overall performance.

**Provenances diversification**

With the development of internet, global social media platforms have grown up. Until the beginning of February 2020, most of COVID-19 infected cases were in China. In addition to daily life related local information, lots of overseas elements concerning COVID-19 topics were popular on Chinese social media platforms. As it is shown in Figure 7, 16.25% of the infodemic during this period deals with foreign topics and half of it is accompanied with elements in a foreign language. Due to the culture gap and the language barriers, it is difficult for the public to defuse misinformation in this way. Even worse, a few topics are attached with elements in non-English foreign languages. As a Public Health Emergency of International Concern, COVID-19 related information on foreign subjects may become sensitive topics in international relations. Therefore, it is important to detect these infomedic in time and mitigate its negative impact.

**Selective review**

A dataset is a collection of instances. When working with neural network approaches, a few datasets are required for different purposes. Owing to the development of big data and artificial intelligence, there have been a great number of datasets for various applications. However, these datasets cannot be used directly

---

**Figure 4.** The word cloud generated by the keywords of Chinese infodemic from the end of January to the beginning of February in the WeChat mini app “Jiaozhen”.

**Figure 5.** The 10 most frequent used keywords of Chinese infodemic from the end of January to the beginning of February in the WeChat mini app “Jiaozhen”.

---

(Luo et al. 2023)
for infodemic detection during the COVID-19 outbreak and limit the effectiveness of neural network approaches in this implementation. Therefore, a selective review of existing datasets is synthesized based on the 4P framework from four perspectives, focusing on pointing out their drawbacks when being utilized for infodemic detection.

**Domain perspective**

Infodemic is a portmanteau between information and epidemic and it has a deep connection with medical science. Therefore, representative datasets from the medical domain, particularly datasets about the COVID-19, are considered in this sub-section as shown in Table 1. Since COVID-19 was first identified in China, China National Center for Bioinformation has firstly created an official website for gathering local and international 2019 Novel Coronavirus Resource (2019nCoVR). Open source datasets about 2019-nCoV Sequences, COVID-19 Clinical Records and so on, have made a great contribution to virus rapid detection, medical treatment and vaccine development. In addition to bioinformation, a number of medical datasets are presented in the form of images, such as the computed tomography scan, the magnetic resonance imaging, etc. Generally, these datasets are long term collections of constructed data led by governments along with professional institutes for supporting health professionals or scientific researchers. Although there are some academic terms in infodemic, similar datasets as the ones mentioned above cannot be considered as sources for infodemic detection datasets due to their professionalism.

In contrast to professional medical datasets for scientific research, general medical datasets may have a higher potential to be used for infodemic detection. These datasets are usually made of uncontracted data, such as patients’ questions or sentences extracted from articles. Despite prevention is always better than cure, people pay less attention to it in daily life. As a result, context in general medical datasets consider more about diagnosis and treatment rather than prevention, which does not fit well the circumstances when a pandemic spreads.

**Event perspective**

COVID-19 related topics have caught tremendous attention during its outbreak. Therefore, representative
datasets dealing with events related to COVID-19, particularly events about infodemic detection, are detailed in this sub-section as shown in Table 2. With more and more concerns about the rapid spreading of misinformation, some projects have been launched to tackle this phenomenon. COVIEWED\textsuperscript{16} is a typical one who plans to release a browser widget that highlights potential true/false claims on a web page to assist users in their information gathering process. This project is still at a preliminary stage where thousands of links to COVID-19 related articles are collected and stored in TSV files. As one popular Chinese online community for health professionals, DXY.cn works similarly as WeChat mini app “Jiaozhen” by labeling hot online articles as true, false or questionable since the end of January 2020 in which fake articles account for 80% approximately. These labeled data\textsuperscript{17} should be involved as important components when building infodemic detection datasets. However, it is not an ideal way to use them directly to train neural network methods due to its limited amount and imbalanced classes.

Because of quarantine measures caused by the COVID-19, events such as food delivery, online shopping and interactions on social media have received more attention than usual. No doubt, datasets about take-away deliverymen records from Ele.me\textsuperscript{18}, E-commerce records from AliExpress\textsuperscript{19} and microblog records from Sina Weibo\textsuperscript{20} can also be used to mitigate the negative impact of the COVID-19 on the society and economy. Although the above-mentioned data are not directly related to infodemic detection, some of them can be adapted to enlarge the datasets. Since these events are closely related to the COVID-19, the involved information may be considered as a good source for exposing a more balanced perspective on the classes by gathering additional true or questionable examples.

**Display perspective**

Text is the main form of infodemic during the COVID-19 outbreak while images have the highest basic reproduction number. Because of the great deal between texts and languages, datasets about text processing will be discussed in the next subsection from the language perspective. Representative datasets involved in this subsection emphasize datasets about image processing, particularly image datasets for text recognition as almost all COVID-19 infodemic images are embedded with texts. As presented in Table 3, there are already widely used datasets for detecting and recognizing

| Table 1. Representative datasets from the domain perspective. |
|-----------------------------------------------|
| Datasets names | Datasets descriptions |
| Datasets related to COVID-19 | 2919-nCoV sequences\textsuperscript{12} | By 15th September 2020, it contains 105,133 2019-nCoV sequences from 1134 locations. |
| | COVID-19 clinical records\textsuperscript{12} | By 15th September 2020, it contains 208 COVID-19 patients’ clinical records, including onset dates, clinic tests, etc. |
| Datasets non-related to COVID-19 | TCGA-COAD\textsuperscript{13} | It contains 461 cases’ radiology data from The Cancer Genome Atlas Colon Adenocarcinoma collection. |
| | Medical semantic textual similarity\textsuperscript{14} | It contains thousands of paired health-related questions covering 10 respiratory diseases, such pneumonia, asthma, etc. |
| | Medical information extraction\textsuperscript{15} | It contains 3,984 medical sentences extracted from PubMed abstracts. Relationships between discrete medical terms are annotated. |

| Table 2. Representative datasets from the event perspective. |
|-----------------------------------------------|
| Datasets names | Datasets descriptions |
| Datasets related to infodemic detection | COVIEWED COVID-19 news corpus\textsuperscript{16} | It contains thousands of links to COVID-19 related articles in TSV files. |
| | DXY rumours\textsuperscript{17} | It contains 285 hot online articles labeled as true, false or questionable. |
| Datasets non-related to infodemic detection | Take-away deliverymen records\textsuperscript{18} | It contains data from February 2020 in 4 parts classified as order, distance, deliverymen and action. |
| | E-commerce records\textsuperscript{19} | It contains data from a mature market and a new market in two parts classified as items and buyers. |
| | Microblog records\textsuperscript{20} | It contains half-year sample data including microblog content, forward count, comment count, like count. |
Chinese characters in natural scenes\textsuperscript{21} or handwritten scenes.\textsuperscript{22} Moreover, lots of research has been carried on datasets of hand-written digits.\textsuperscript{23} These datasets are efficient for training learning techniques and pattern recognition methods. However, the collected samples are mainly composed of characters, words and short texts while the embedded infodemic texts are much longer.

As the COVID-19 has become a Public Health Emergency of International Concern, some texts embedded infodemic images are presented as a mixture between Chinese and foreign languages, where the former is used as a caption for foreign languages. Meanwhile, certain infodemic images with texts in natural scenes are mixed with personal comments generated by image caption apps. Therefore, hybrid datasets of multi-languages\textsuperscript{24} and multi-scenes\textsuperscript{25} should also be taken into account when the infodemic dataset is constructed.

**Language perspective**

Text information accounts for 73\% of infodemic at the initial stage of the COVID-19 outbreak. Moreover, text embedded images will be converted to text information after text recognition processing. Therefore, the work to understand text information plays a quite important role for infodemic detection. As texts are always written in one or several languages, representative datasets discussed in this subsection are divided into two groups as shown in Table 4. The procedure for infodemic detection can be considered similar as text classification, classifying social media texts during outbreaks as rumors or non-rumors. As the main source for Chinese rumors detection dataset,\textsuperscript{2} Sina Weibo encourages users to report possible fake tweets. Afterwards, the reported tweets are scrutinized and judged by a group of elite users where verified fake tweets will be labeled. Due to the difference between general rumors and infodemic, raw data from Sina Weibo cannot be used directly to train neural network methods for infodemic detection. In contrast, datasets about named-entity recognition\textsuperscript{26} and automatic summarization\textsuperscript{27} should be fully utilized to assist manual work, particularly for questionable information who requires elaborate examination.

Although most of the information shared on Chinese social media is in Chinese, 8\% of collected infodemic is displayed with elements in foreign languages.
languages. Moreover, when a piece of information about another country is written in Chinese, it is necessary to check the domestic news in its local language to verify how reliable it is. Therefore, parallel text datasets for machine translation should not be ignored. Besides, it is better to label each sentence with the language it is written in, when a dataset deals with multiple languages translations. This is an efficient way to defuse made up rumors, just as one widely shared piece of information, written in Spanish, was made up as one in Italian.

Research directions

Although Wuhan, the first identified COVID-19 outbreak city, is unlocked since 8th April 2020, the growth of COVID-19’s spread is not coming to an end. Any rise of infections may lead to the increase of infodemic. A large-scale labeled dataset can enhance the overall performance of infodemic automatic detection when neural network methods are used. Moreover, the occurrence of Public Health Emergency of International Concern is low but its bad impact from alongside infodemic is tremendous. It is critical to take the chance to gather infodemic data and construct large-scale datasets which can be used directly for infodemic automatic detection when similar outbreaks happen. Therefore, research directions, including recommendations, about constructing a large-scale dataset for Chinese infodemic automatic detection are proposed.

Prevention attention from a domain perspective

A total of 43% of the infodemic concerns prevention measures. Therefore, prevention measures against pandemics should be considered as the essential part of infodemic datasets. The sentences involved are required to be as simple as the ones for general medical datasets, while their goal need to be shifted from diagnosis and treatment to prevention. Some handbooks issued by professional institutions and advice from the World Health Organization website should be taken as good resources. Although professional medical datasets are made for supporting health professionals or scientific researchers, some elements are widely misused in infodemic. Therefore, extracting frequently used elements, particularly some academic terms, and using them in easily understandable content should be considered when constructing infodemic datasets.

Problem orientation from an event perspective

“COVID-19” related topics have received tremendous public attention during the outbreak of this disease. Although it is difficult to predict the next one, some epidemic-orientated infodemic datasets should be prepared in advance for the sake of prevention. Public health events referred in the Regulation on the Urgent Handling of Public Health Emergencies launched by the Chinese State Council on May 7th, 2003 can be used as a case in point. According to the nature of the event, the degree of negative impact and the spreading scope, these events are ranked in four levels from the significant level to the general level. Data concerning the events ranked at the significant level should be collected when the infodemic dataset is built. Moreover, data from artificial scenarios can also be utilized to enlarge the total amount, if there is not enough data from real scenarios.

Patterns interaction from a display perspective

A total of 73% of the infodemic is displayed as text information while the basic reproduction number of the photo sharing social media platform during the COVID-19 outbreak is as high as 2.25. Therefore, it is critical for infodemic datasets to gather both text and image information. In contrast to images used for character detection and recognition from natural scenes or hand-written scenes, images working with optical character recognition techniques are supposed to have better performance. In this case, the converting speed may become a new challenge when the processed volume is high. Since the total amount and the basic reproduction number of video infodemic is relatively low, human detection can work efficiently to handle this process. However, some videos with various formats should still be included inside infodemic datasets, where neural network methods can be trained to figure the media type and pass them out for further manual detection.

Provenance diversification from a language perspective

Either for texts or text embedded images, understanding text information and classifying them into different types are essential techniques for infodemic detection. Differently from the traditional way that classifies social media texts as rumors or non-rumors, data in infodemic datasets should be annotated into three groups: true, fake and questionable as with the WeChat mini app “Jiaozhen”. Besides, data from artificial scenarios may be utilized to expose a balanced perspective among the three groups. Due to the uncertainty during public health emergencies, manual detection is considered as a mandatory step to deal with data which are classified as questionable. To facilitate manual work, these data are suggested to be attached with named entity recognition tags and paired short summaries. Moreover, either presenting only in Chinese or displaying with elements in foreign languages, information about another country is supposed to be annotated with the name of its geographic origin. The attached named entity recognition tags and paired short summaries are also required to have a translation in the source languages.
Conclusions and future works

In this paper, the first framework of Chinese infodemic was presented, which consisted of four features: Prevention Attention, Problem Orientation, Patterns Interaction and Points Globalization. This framework was built on deep analysis of hot ambiguous topics labeled as false and questionable in the WeChat mini app “Jiaozhen” from the end of January to the beginning of February. Afterwards, a selective review of 20 representative datasets is synthesized based on the 4P framework from four perspectives, focusing on pointing out their drawbacks when being utilized for infodemic detection. To facilitate infodemic automatic detection when similar outbreaks might happen in the future, recommendations about constructing a large-scale dataset for Chinese infodemic automatic detection along the layers of the 4P framework were proposed at the end.

Three main areas that deserve further study are identified. The first issue is to build a Chinese infodemic dataset in terms of the 4P framework, particularly considering opinions proposed in Section 4 Research Directions. A second line of interest is to take widely used recurrent neural networks or convolutional neural networks to test the effectiveness of the dataset. As preliminary works8 have attempted to train BERT (Bidirectional Encoder Representations from Transformers) on general rumor datasets for infodemic automatic detection, finally we would like also to train BERT on the new built dataset and compare its overall performance with previous works.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work is supported in part by the Japan Society for the Promotion of Science.

ORCID IDs

Jia Luo https://orcid.org/0000-0003-3914-8865
Rui Xue https://orcid.org/0000-0001-8189-6504
Jinglu Hu https://orcid.org/0000-0002-5601-7261

References

1. Zarocostas J. How to fight an infodemic. *Lancet* 2020; 395(10225): 676.
2. Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks. In: *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, New York, 9–15 July 2016, pp.3818–3824. New York: AAAI Press.
3. Yu F, Liu Q, Wu S, et al. A convolutional approach for misinformation identification. In: *Proceedings of the 26th international joint conference on artificial intelligence*, Melbourne, Australia, 19–25 August 2017, pp.3901–3907. California: IJCAI.
4. Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: *Proceedings of the 25th ACM international conference on multimedia*, Mountain View, CA, 23–27 October, pp.795–816. New York: ACM.
5. Wang Y, Ma F, Jin Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, London, July 2018, pp.849–857. New York: ACM.
6. Yu F, Liu Q, Wu S, et al. Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Comput Secur* 2019; 83: 106–121.
7. Wu K, Yang S and Zhu KQ. False rumors detection on Sina Weibo by propagation structures. In: *2015 IEEE 31st international conference on data engineering*, Seoul, South Korea, 13–17 April 2015, pp.651–662. New York: IEEE.
8. https://github.com/lqhou/NLP_Bus_Pro/tree/master/fake_news
9. http://www.sic.gov.cn/Column/595/0.htm
10. Kulev I, Janevska E, Jovanovik M, et al. Text classification using semantic networks. In: *8th International conference for informatics and information technology*, Pune, India, 7–8 November 2011. Berlin: Springer.
11. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The Covid-19 social media infodemic. *Sci Rep* 2020; 10: 16598.
12. https://bigd.big.ac.cn/ncov?spm=5176.12282016.0.0.602411d2f7npvn#progress
13. Kirk S, Lee Y, Sadow CA, et al. *Radiology data from The Cancer Genome Atlas Colon Adenocarcinoma [TCGA-COAD] collection*. The Cancer Imaging Archive, 2016. Maryland: Bethesda.
14. https://tianchi.aliyun.com/competition/entrance/231776/introduction
15. Dumitrache A, Aroyo L and Welty C. *CrowdTruth measures for language ambiguity*. In: *Proceeding of the LD4IE workshop*, ISWC, 2015. 12 October 2015, Bethlehem, Pennsylvania, USA.
16. https://www.coviewed.org/
17. https://github.com/BlankerL/DXY-COVID-19-Data/blob/master/README.en.md
18. https://tianchi.aliyun.com/competition/entrance/231777/introduction?spm=5176.12281949.1003.1.493e4c2afPnb3R
19. https://tianchi.aliyun.com/competition/entrance/231718/introduction
20. https://tianchi.aliyun.com/competition/entrance/231574/information
21. Yuan TL, Zhu Z, Xu K, et al. A large Chinese text dataset in the wild. *J Comput Sci Technol* 2019; 34(3): 509–521.
22. Zhang H, Guo J, Chen G, et al. *HCL2000-A large-scale handwritten Chinese character database for handwritten character recognition*. In: *2009 10th international conference on document analysis and recognition*, Barcelona, 26–29 July 2009, pp.286–290. New York: IEEE.
23. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86(11): 2278–2324.
24. Yao C, Bai X, Liu W, et al. Detecting texts of arbitrary orientations in natural images. In: 2012 IEEE conference on computer vision and pattern recognition, Providence, RI, 16–21 June 2012, pp.1083–1090. New York: IEEE.

25. De Campos TE, Babu BR and Varma M. Character recognition in natural images. VISAPP 2009: 2: 7.

26. He H and Sun X. F-score driven max margin neural network for named entity recognition in Chinese social media. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers, Valencia, 3–7 April 2017, pp.713–718. Stroudsburg, PA: Association for Computational Linguistics.

27. Hu B, Chen Q, and Zhu F. (2015, September). LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September. pp. 1967–1972. Pennsylvania: Association for Computational Linguistics.

28. Tian L, Wong DF, Chao LS, et al. UM-corpus: a large English-Chinese parallel corpus for statistical machine translation. In: Proceedings of the ninth international conference on language resources and evaluation (LREC’14), Reykjavik, Paris, France, May 2014, pp.1837–1842. European Language Resources Association (ELRA).

29. Eisele A, Chen Y and Multi UN. A multilingual corpus from United Nation documents. In Proceedings of the seventh international conference on language resources and evaluation (LREC’10), Paris, France, 17–23 May 2010, pp.924–929. European Language Resources Association (ELRA).

30. https://new.qq.com/omn/20200306/20200306A06CWS00.html

31. Liang T. Handbook of COVID-19 Prevention and Treatment: Compiled According to Clinical Experience. The First Affiliated Hospital, Zhejiang University School of Medicine, 2020. Hangzhou, China.

32. http://www.gov.cn/zwgk/2005-05/20/content_145.htm