Computer Science and Information Systems 19(3):1115–1132 https://doi.org/10.2298/CSIS210822014S

COVID-19 Datasets: A Brief Overview

Ke Sun[1], Wuyang Li[1], Vidya Saikrishna[2], Mehmood Chadhar[3], and Feng Xia[3,⋆]

[1] School of Software, Dalian University of Technology,
Dalian 116620, China
{kern.sun,wuyang_li}@outlook.com
[2] Global Professional School, Federation University,
Ballarat 3353, Australia
v.saikrishna@federation.edu.au
[3] Institute of Innovation, Science and Sustainability, Federation University Australia,
Ballarat 3353, Australia
m.chadhar@federation.edu.au
f.xia@ieee.org

Abstract. The outbreak of the COVID-19 pandemic affects lives and social-economic development around the world. The affecting of the pandemic has motivated researchers from different domains to find effective solutions to diagnose, prevent, and estimate the pandemic and relieve its adverse effects. Numerous COVID-19 datasets are built from these studies and are available to the public. These datasets can be used for disease diagnosis and case prediction, speeding up solving problems caused by the pandemic. To meet the needs of researchers to understand various COVID-19 datasets, we examine and provide an overview of them. We organise the majority of these datasets into three categories based on the category of applications, i.e., time-series, knowledge base, and media-based datasets. Organising COVID-19 datasets into appropriate categories can help researchers hold their focus on methodology rather than the datasets. In addition, applications and COVID-19 datasets suffer from a series of problems, such as privacy and quality. We discuss these issues as well as potentials of COVID-19 datasets.

Keywords: COVID-19, Data science, Datasets, Artificial intelligence.

1. Introduction

In late 2019, a novel virus, named COVID-19 emerged all over the world. This virus was declared as a global pandemic by the World Health Organization on March 11, 2020. The COVID-19 has incalculable influences on the world's health, social and economic conditions [54]. With the increase in the number of people infected with COVID-19 every day, it is essential to find a fast and effective way to manage the problems caused by the COVID-19 for biological, medical, and public health issues. Recently, effective utilisation of Artificial Intelligence (AI) technologies [33, 50] to perform analysis, prediction and diagnosis are ongoing researches to fight against coronavirus [2, 47, 59]. We know that AI-based models rely on the available datasets. Thus, datasets play a key role in fighting against the COVID-19 pandemic [34, 37, 79].

⋆ Corresponding author

With the advancement of the Internet and digital media technologies, many open datasets are now available on the websites of research institutions [39]. These datasets are public and can be downloaded for free. For conveniently use of research, this paper provides a summary of the datasets collected from official websites and academic publications. The purpose of our work is to provide researchers, professionals and scholars a quick reference of datasets in application. Based on usages, datasets are organised into several meaningful taxonomies such as, dataset fields, organization structure, and purpose. The categories identified for COVID-19 datasets include: time-series, knowledge and media datasets. We briefly introduce each category of datasets in the following.

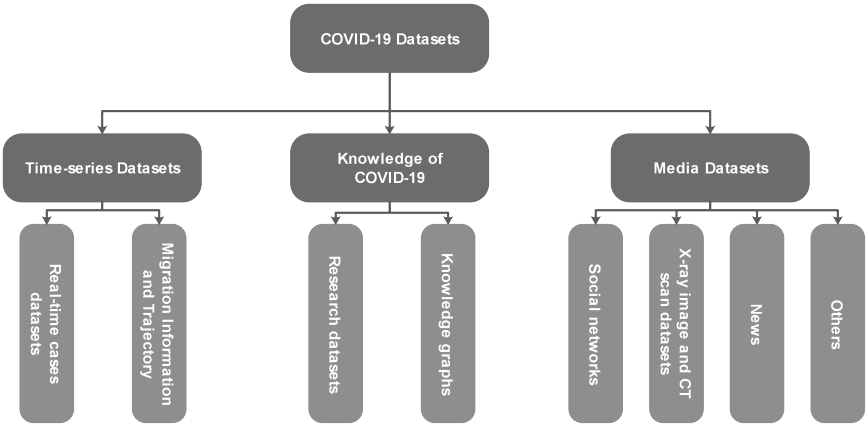

Fig. 1. Taxonomy of COVID-19 open-source datasets

We review the COVID-19 datasets according to the data format and applications. The first category of COVID-19 datasets is the time-series datasets [78]. This kind of datasets is relatively simple in structure and usually contains time information, death statistics, population movement data, etc. Thus, they can be used without complicated preprocessing, such as directly showing the number of cases every day. In addition, this kind of datasets can be used to predict the impact of the disease on human society [6] and trend of virus growth in the future [38].

The second category of datasets is knowledge base of COVID-19, for instance, scholar knowledge graph [36], medical knowledge graph [73]. A part of COVID-19 knowledge bases [46] contain complex contents in datasets, such as text data and image data. Therefore, it is necessary to extract and preprocess the contents to structured data before using the inner information, such as converting the text data into knowledge graphs. In addition, the COVID-19 knowledge base [46] has location information with timestamps and can be used to estimate the risk of infection on people.

The third category of datasets is the media datasets. They have many types of information such as text, image and video. There are lots of media datasets of COVID-19 published on the Internet such as, social networks [8], news information [52]. These information can be used to analyse the current emotional state of people and can be used to

monitor current concerned topics for the public. The taxonomy of COVID-19 datasets are summarized in Fig 1.

The rest of this paper is organized as follows. The time-series datasets are introduced in Section 2. The knowledge base are described in Section 3 and the media datasets are presented in Section 4. In Section 5, several issues of the datasets are discussed followed by conclusions in Section 6.

2. Time-series Datasets of COVID-19

This section is organized to cover information on the source of real-time case datasets and migration datasets around the COVID-19 time. The comparison of several representative datasets is presented in Table 1.

Table 1. Comparison of COVID-19 Time-series Datasets

Dataset	Data type	Level	Main Contents
nCoV2019 [72]	Text, values	Country, province	Symptoms, key dates, and travel history
COVID-19 [13]	Text, values	Country, state, city	Daily case reports, regional coordinates
covid19india-cluster	Text, values	India	Outbreak and transmission COVID19 in India
CoronaWatchNL	Text, values	World	COVID-19 case number, age and sex information
qianxi[¶]	Text, map	China	Population migration information of China
2019-nCov-data	Text, values	China	Migration, news, and rumor data
OAG[‖]	Text, values	World	Flight schedules data
IATA[**]	Text, values	World	Flight schedules data

2.1. Real-time cases datasets

Real-time cases datasets are import for fighting against the COVID-19 disease. For example, these real-time cases datasets could provide people intuitive trend information of COVID-19 and can support public health decision making. To analyse and track the COVID-19 pandemic from real-time cases, Xu et al. [72] collected the real-time cases data from national health reports, and provincial health reports. They also collected information from online reports to supplement their dataset. The dataset contains richer information compared with other real-time case datasets. For example, it has geo-coded information, such as travel history, symptoms, and key dates information (dates of onset, admission, confirmation). In addition, this dataset is being updated in real-time and can be downloaded for free.

Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) published the Johns Hopkins epidemiological dataset, including daily case reports and time-series summary tables [13]. The dataset has data fields, including country names, state names, place names, the time of last update, and regional longitude. It is available for the public and can be downloaded from the Github repository. Several studies were conducted based on this dataset. For example, Domenico et al. [5] utilized a variant regressive model to predict the epidemiological trend of COVID-19. Punn et al. [48] developed

a deep learning-based model to monitor people's behaviour every day. The model could utilize the real-time dataset of the Johns Hopkins dashboard to predict the trend of the COVID-19 spreading across nations in the future. Tátrai et al. [58] used the same dataset to investigate how well the logistic equation can predict the influence of the COVID-19 pandemic on the place where the outbreak occurred. The proposed model was used to estimate the risky point, the date of reaching a certain percentage of infections and the number of infected persons in the future.

Several official institutions recorded real-time data and published them to the public. For example, the National Health Commission of the People's Republic of China daily publishes the latest cases information COVID-19 on the official website[†]. Roda et al. [51] used the data to predict the cases of COVID-19 in Wuhan city after lockdown. Singapore collected and published the real-time cases' information of COVID-19 on the Ministry of Health official website[‡]. In addition, the Singapore reports cover detailed analysis of the COVID-19 data. The Tianjin Health Commission daily published the local COVID-19 cases in the form of online press release on their official website[§]. The detailed report along with analysis can be obtained from the same official website.

2.2. Migration information and trajectory

Population migration influences on the spread of virus. To track and discover their relations, it is necessary to record the population migration for studying the trend of virus transmission. There are many open datasets to track the COVID-19 transmission shared on the Internet. A well known migration dataset is available on the Baidu Migration site[¶]. This data can be used to study the pattern of population migration during the Spring Festival of China [27,71]. In addition, researchers can utilize this data to visualize population migration around China. In this dataset, "qianxi" index is used to reflect the size of population moving in or out, and the cities can be compared horizontally. The intensity of city travel is calculated as the ratio of the number of people travelling to the city to the resident population in the same city. In addition, Baidu built a data federation platform (Baidu FedCube), which provides usage instructions and data download services.

Migration datasets of confirmed COVID-19 patients usually contain the travel information including start time, end time, travel type, number of trips, travel description, departure station, arrival station, and other information of the confirmed patients. Several studies were conducted based on migration datasets. For instance, to estimate the geographical scope of the spread of COVID-19 and its potential risks, Lai et al. [26] proposed a deep learning-based model, which could learn from population migration dataset and give future prediction results. Huang et al. [20] attempted to utilize the nationwide mobility data to study the economic impact caused by the COVID-19.

Several trajectory data of proprietary airline are commercially available, such as Official Airlines Guide (OAG) database[‖], International Air Travel Authority (IATA) database[**]. The IATA database contains about 90% of passenger information of commercial flights,

[†] https://www.nhc.gov.cn/
[‡] https://www.moh.gov.sg/
[§] https://wsjk.tj.gov.cn/
[¶] https://qianxi.baidu.com/
[‖] https://www.oag.com/
[**] https://www.iata.org

including the direct starting point from Wuhan to the destination and the indirect starting point from Wuhan with a connecting flight to the final destination [26]. Chinazzi et al. [9] utilized a global aggregate population disease transmission model and proprietary airline data to predict the impact of travel restrictions on the spread of the COVID-19. Proprietary airline data can be used to evaluate the capacity to detect the COVID-19 of different locations. For instance, Rene et al. [44] utilized a Bayesian-based model and the proprietary airline data to estimate the capacity of 194 locations. In addition, the authors designed a mathematical model to calculate the rate of local residents being infected by foreign tourists.

Descartes Labs collected and released DL-COVID-19 dataset, which is mobility dataset at state and county level of US [67]. The dataset was published on the GitHub repository under the Creative Commons Attribution license. Based on the DL-COVID-19 dataset, Michael et al. [67] have found significant changes in the flow of people caused by COVID-19 in US and around the world through mobility data in US.

Transportation Security Administration (TSA) published a confirmed COVID-19 cases dataset[††]. They notified the public about the airport locations where the employees belonging to TSA were found positive for the COVID-19 virus. TSA listed airports with confirmed COVID-19 cases and also the corresponding employees inflicted by the virus.

3. Knowledge Bases of COVID-19

This section is divided into two subsections. The first subsection introduces research datasets of COVID-19 from knowledge point of view. The second subsection deals with the knowledge graphs and their importance. Table 2 gives the summary of these datasets.

Table 2. Summary of COVID-19 knowledge bases

Dataset	Type	Size	Main Contents
CORD-19 [63]	Article	128,000 articles	Coronaviruses related
CORD-NER [66]	Text	75 entity types	Entity types related to the COVID-19
COVID-19-epidemiology[‡‡]	Knowledge graph	374 instances	Epidemiological knowledge
covid19kg [12]	Knowledge graph	4016 nodes, 10 entity types	Virus protein, potential drug target, etc
covid-19-medical[‡‡]	Knowledge graph	383 instances	Clinical knowledge

3.1. Research dataset for COVID-19

Since the outbreak of COVID-19, many research papers for the study and analysis of the new coronavirus have surged, especially in the fields of medicine and biology. CORD-19 [63] is one of a extensive machine-readable and large new coronavirus paper collection for data mining to date, which contains historical and the latest scientific research papers of the coronavirus. The CORD-19 was provided by the leading research groups of Semantic Scholars at Allen AI. The dataset has a collection of more than 128,000 academic

[††] https://www.tsa.gov/

articles and 59,000 full texts on new coronaviruses, containing articles related to such as, COVID-19, SARS (Severe Acute Respiratory Syndrome), and MERS (Middle East Respiratory Syndrome). The dataset contains more than 50,000 metadata files of coronavirus research articles, including but not limited to COVID-19.

CORD-19 is an open knowledge research dataset, and it is free for use by the global research community. For instance, the worldwide AI research community utilized this dataset and data mining methods to fight against the COVID-19. Since its release, CORD-19 has been downloaded more than 75,000 times. It becomes the basis of many COVID-19 text mining and discovery systems and could promote generating new insights into the fight against the ongoing COVID-19. In addition, the dataset has links to other publication databases such as PubMed, Microsoft Academic Graph, Semantic Scholar, and WHO through unique keywords. Thus, CORD-19 has richer information than other knowledge bases.

At present, CORD-19 has been used in information extraction, information retrieval and knowledge graphs by natural language processing and deep learning techniques. Besides, it can also be used in multiple directions, such as question answering [43], pre-trained language models [31], summarization [55], and recommendations [53]. To inspire developers to find new insights in the large-scale COVID-19 epidemic, Kaggle utilized the CORD-19 dataset to host an open research dataset challenge. The research challenge includes the topic of tracing the history of the virus [35], the study of the transmission characteristics of the virus, the diagnosis of the virus, etc.

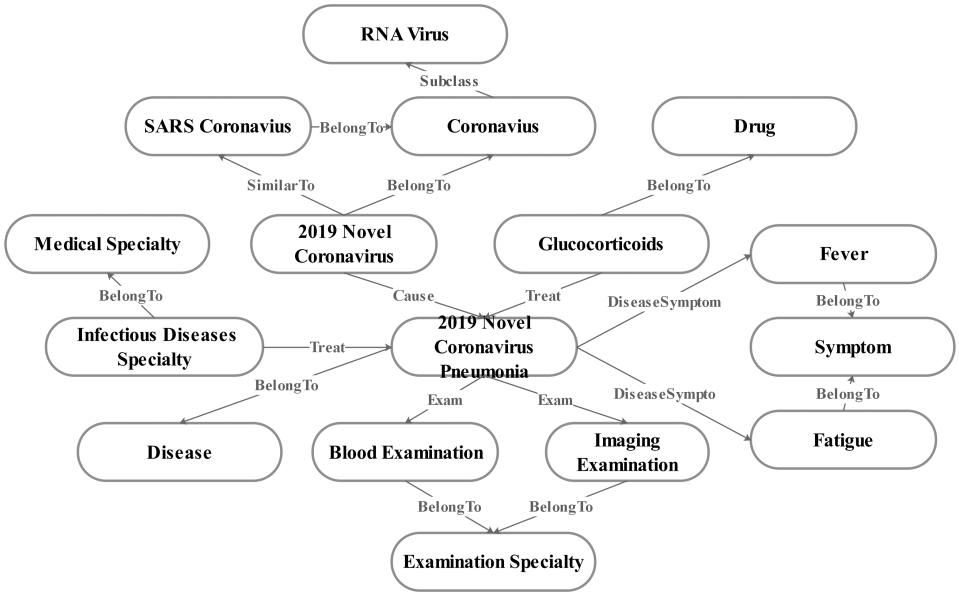

Fig. 2. A medical knowledge graph of COVID-19. This knowledge graph contains entities such as viruses, bacteria, epidemics, and infectious diseases. Entities connected to each other with relations, such as “Subclass”, “SimilarTo”, and “Cause”

3.2. COVID-19 knowledge graphs

Knowledge graphs associated with COVID-19 attracts lots of attention of researchers. Many kinds of knowledge graphs are published such as, pathophysiology knowledge graph [12] and biomedical knowledge graphs [1, 81]. Knowledge graphs are useful in a wide range of applications, for instance, knowledge answer system of COVID-19 [28], auxiliary diagnostic system of COVID-19 [30].

Directly getting desired knowledge from lots of research articles is relative time-consuming. An efficient approach is to obtain knowledge from pre-constructed knowledge graphs. Knowledge graphs utilize topology to integrate data and cover a wide range of knowledge. For instance, the COVID-19 knowledge graphs contains biological processes, drug-target interactions, genes and proteins of the new coronavirus [16]. Thus, the COVID-19 knowledge graph can even help researchers discover hidden interactions of protein.

Metrological analysis and visualization of knowledge graph methods can help extract and formalize structured knowledge. For example, network embedding theories [19, 56] and graphic visualization technologies can be used to visualise knowledge graphs. A visualization of a medical knowledge of COVID-19[‡‡] is presented in Fig 2. The visualization of knowledge graphs can provide intuitive connections between entities and can promote researchers to better understand knowledge in less amount of time. Moreover, in the visual web application of knowledge graphs, users can browse and query the network, filter nodes or edges according to their own needs, or calculate the path between nodes of interest.

There are several knowledge graphs built from CORD-19 [63]. The literature of medicine and biology are the main contents of this dataset. Several natural language processing methods were applied to this dataset to construct knowledge graphs. The constructed knowledge graph contains information on medicine and biology, which are important for researchers [81]. In addition, the rich relations in knowledge graphs could help researchers to discover hidden information and thus contribute to fight against COVID-19. To apply CORD-19 to the field of knowledge graphs, the first work is to create a named entity recognition (NER) dataset. Wang et al. [66] built an entity recognition dataset for CORD-19. The NER dataset contains 4 different sources. The entities in the NER dataset are mapped into 75 fine-grained entity types. The results obtained by identifying named entities on the CORD-19 dataset helps in constructing knowledge graphs. For example, it can be applied to build medical knowledge graphs.

Several other researchers devoted themselves to constructing knowledge graphs of COVID-19. Domingo-Fernández et al. [12] constructed an extensive a knowledge graph based on the COVID-19 paper collection. The knowledge graph contains information of the COVID-19 virus protein, potential drug target and the biological transmission path of the virus. The knowledge graph could provide a new research perspective for exploring the physiology of COVID-19 cases. First, the authors filtered the unimportant information from the available source. Second, they collected a part of free and open scientific articles related to COVID-19. Then, the collected articles were scored and ranked by using modelling language tools based on importance. Finally, the knowledge graph was constructed from the selected articles.

[‡‡] https://openkg.cn/

The integration of large-scale knowledge graphs and information mining functions are urgently needed for filtering plenty of new coronaviruses, especially in the field of medicine. The drug knowledge graphs can help medical researchers quickly find potential drug candidates. For instance, Ge et al. [16] designed a knowledge graph building method, which is a data-driven drug framework. The built knowledge graph is virus related, including knowledge of drug-target, protein-protein interactions. Three different types of nodes exist in the knowledge graph, namely drugs, human targets and viral targets. Entities in knowledge graph are connected with edges, which describe the relationship, similarities and interaction between entities. A total of seven networks are considered to construct the knowledge graph, including human protein-protein interaction network, human target-drug interaction network, and so on.

Network embedding algorithms [70] are well known for network analysis. It can be applied to knowledge graphs to predict the drug candidate list, saving the time and cost of discovering effective drugs for disease. For instance, Hong et al. [18] proposed a relation extraction method based on deep learning technology, namely BERE. The method can be applied to mining large-scale literature. Relying on this method, only a small number of candidate drugs on the list need to be manually checked, thus, the list of candidate drugs is further narrowed.

4. COVID-19 Related Media Datasets

This section covers datasets collected from social networks, news and other media sources in three different sub-sections. These datasets are summarized in Table 3.

Table 3. Summary of COVID-19 related media Datasets

Dataset	Format	Contents
COVID19socialscience [75]	Text	Tweet of 69 institutional/media Twitter accounts
covid19twitterevent [83]	Text, JSON	COVID-19 Events from Twitter
covid19twitter [4]	Text, JSON	Twitter chatter of COVID-19
CoronaVis [22]	Text, JSON	Personal opinions, facts, news, status
COVID-19-TweetIDs [8]	Text	50 million tweets
COVID-19-InstaPostIDs [77]	Text	Public posts from Instagram
covid19_dataset [15]	TSV	Tweet, user ID and Weibo ID
COVID-CT [82]	Text, xlsx,image	CT scans from medRxiv, bioRxiv, etc
covid-chestxray-dataset [10]	Text, csv, image	Chest images of COVID-19 or other pneumonias
COVID-19 [45]	Image	Normal, pneumonia chest images

4.1. Social networks

Many datasets for social networks [69] are published such as, Twitter and Facebook datasets. These datasets can be used to support urgent research to address the outbreaks caused by COVID-19. Considering that there is no specialized collection of tweets posted

by the government or news media, Yu [75] published a COVID-19 Twitter dataset for social science research, which is built on the keywords of coronavirus and COVID-19.

Department of Social Psychology, Universitat Autònoma de Barcelona published an Institutional and News Media Tweet dataset of COVID-19 for social science research [75]. The dataset was obtained from Twitter accounts of 69 institutions/news media, including 17 government and international organizations and 52 news media in North America, Europe, etc. There are 8 categories in the collection: "Government Tweets" (government, international agencies, etc.), "US News Tweets", "British News Tweets", "Spanish News Tweets", "German News Tweets", "France News Tweets", "China News Tweets", and "Additional News Tweets". Each category contains different collection targets. This microblogging data can support sociologists to analyse the impact of the pandemic on public interest, health information, and social response to policy-makers [14].

The Department of Computer Science at the University of Missouri published the coronaVirus Twitter (focused on the United States) dataset [22]. They collected and processed more than 100 million tweets related to the novel coronavirus using Twitter Streaming API and Tweepy since March 5, 2020. The collected raw data around 700GB up until April 24, 2020, and saved these collected data in the format of JSON. To improve the usability of data, the dataset has been dynamically processed in real-time, which is stored and being updated in the Github repository. Every single file in dataset contains intra-day data. Date is set as the single file name. The file in the dataset contains 6 different attributes (tweet_id, created_at, loc, text, user_id, verified). The tweet_id represents the unique id of a tweet. The created_at represents the creation time of a tweet. The loc represents the state user location. The text represents the text of the tweet being processed, with all the text in lowercase, non-English characters, and some stop words removed. The user_id means that the exact user name of the pseudo-user ID is converted to an anonymous ID to protect the user's privacy. The verified field indicates whether the tweet is verified (1 or 0), 0 means unverified, 1 means verified. During the pandemic, people were isolated at home, but social media allowed people around the world to stay connected. Collecting information of people shared on social media, such as personal opinions, status and location, can help researchers understand public behaviour during a pandemic. This dataset can be used for such as, sentiment analysis [41, 42], behavioural decision-making [76].

Another public Twitter dataset [8] related to coronavirus was collected by the Information Sciences Institute, University of Southern California. The dataset has more than 50 million tweets from the inception until March 16, 2020, about 450 GB of raw data. The dataset could be used to track scientific coronavirus misinformation and unverified rumours, and help researchers to understand the fear and panic of the public [80]. There is another Twitter dataset of COVID-19 for scientific research [**?**]. This open dataset enables researchers to carry out research projects on emotional and psychological responses to social distance measurements, identification of false sources, and stratified measurements of pandemic emotions.

The first Instagram dataset for COVID-19 was collected by researchers at Queen Mary University of London, England [77]. The dataset is published on a Github repertory. The dataset contains four main parts: (1) publisher information content, (2) post content, (3) like features, and (4) comment metrics content. The posts content part has key attributes, such as captions, hashtag lists, images/videos, likes, comments, locations, dates, and tagged lists. Posters can be public accounts (or public Instagram pages) and datasets

contains information about individuals, fan pages, news agencies, influencers, bloggers, and so on. Each post receives a response, such as a comment issued by a viewer/follower. The dataset helps researchers study the analysis of fake news, false alarms, rumours, the robot population and robot-generated content, and behavioural changes during the spreading of the COVID-19 pandemic.

Georgetown University built a Tweets dataset of COVID-19 using Twitter's Streaming API (Twitter Streaming API) [15]. Tweets related to COVID-19 are defined as tweets with 16 tags, such as 2019nCoV Corona SARI. The dataset has 2,792,513 tweets, 456,878 quotes, and 18,168,161 retweets. Most tweets in dataset were in English (57.1%), followed by tweets in Spanish (11.6%). The dataset was divided into two parts: one was grouped according to the location information in the tweet's content and the other was grouped according to the location information based on the time of the tweet. More than 351,000 tweets in the data have links to news organizations, which accounted for about a tenth of the original tweets of the samples. Researchers found that more than 63,000 tweets were linked to high-quality sources and more than 1,000 were linked to low-quality sources. Currently, the dataset has been used to find a correlation between the virus outbreak and the activity level of local social media. Although rumours and low-quality information still exist, they have little impact on general trends such as the direction of public opinion. The dataset can also be used by natural language processing models for more sophisticated spatiotemporal analysis of information flows and the spread of COVID-19, aiming at identifying rumours and topics [61].

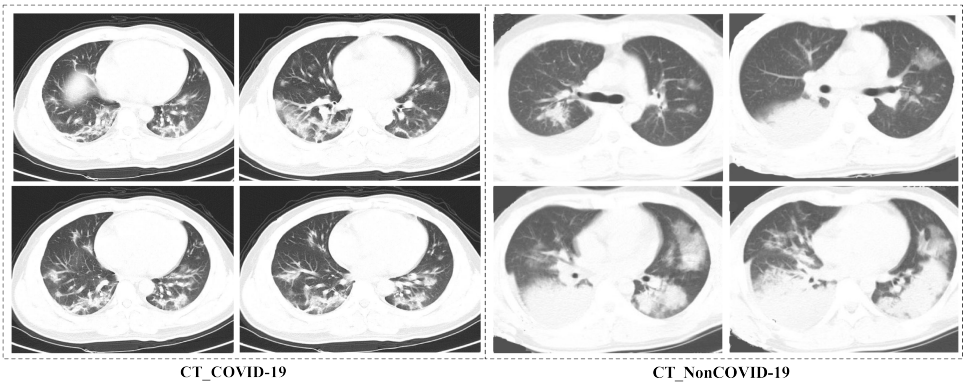

Fig. 3. Examples of CT scans for COVID-19. The left images are the CT scans of a patient's chest infected with COVID-19. The right images are the CT scans of the chest of a normal person

4.2. X-ray image and CT scan datasets

Medical images are important sources to diagnose COVID-19, such as chest CT scans and X-ray images. To improve the efficiency of medical diagnosis, many institutions attempt to automatically diagnosis this disease by exploiting deep learning algorithms and

COVID-19 medical image datasets. A part of institutions has shared the collected COVID-19 medical image datasets on the Internet to promote the research of artificial intelligence in pneumonia diagnosis. In the following, we present these datasets in this part.

Many medical images of COVID-19 can be obtained from GitHub repository or the official website. For example, the COVID_CT dataset [82] was published on GitHub repository. Examples of CT scans for COVID-19 of this dataset are shown in Fig 3. This dataset was collected from COVID-19 related papers published in several open databases such as, medRxiv, bioRxiv. The authors collected images from papers by matching the pneumonia name of image titles. This dataset contains normal and infected CT scans. The subset of COVID_CT contains 349 images from the clinical findings of 216 patients. The two categories of images are stored separately in two different files. Cohen et al. [10] collected open sources and diagnostic data from hospitals and created a public image dataset of COVID-19. This dataset contains chest images of COVID-19 or other types of pneumonia (MERS, SARS and ARDS) of positive or suspected patients. The dataset contains COVID-19 chest X-rays images of 412 people from 26 countries/regions, including 679 images. In addition, this dataset contains clinical records of patients such as blood tests, ICU stay. Ozturk et al. [45] created an integrated COVID-19 X-ray dataset from a part of two open-source datasets, including the dataset published by Cohen JP [10] and the ChestX-ray8 dataset provided by Wang et al. [65]. The integrated dataset covers normal, pneumonia chest images and COVID-19 images of people, containing diagnostic images of 43 female and 82 male patients, and a total of 127 COVID-19 chest images. In addition, the dataset contains the age information of 26 patients. Wang et al. [62] created the largest publicly accessible COVID-19 chest X-ray image dataset, namely COVIDx. This dataset was collected from five open-access data repositories and contains a total of 13,725 patient cases and 13,800 chest images. Several studies [21, 60, 68] have conducted AI-based diagnosis with the help of COVIDx. The authors [64] collected a total of 259 chest CT scans from several hospitals in China and 15 recruited patients. The dataset contains 180 typical viral pneumonia and 79 confirmed SARS-COV-2 cases. This dataset can be obtained from the supplementary data of the study.

4.3. News

News media data serves as a convenient and direct data for the public [74]. People can obtain the exact situation of the pandemic in the current region, and take corresponding measures to protect themselves. We can use the existing data to predict the follow-up development of the epidemic in each region. For instance, researchers can use this kind of dataset to apply statistical analysis to compare pandemic features among different countries, and try to find features that can bring new insights to fight the pandemic [17]. The dataset can also be used with other kinds of datasets, such as regional and subregional social demographic [24].

4.4. Others

A question-and-answer system was developed to obtain the CovidQA dataset [40] to help the research community find answers and gain insight into coronavirus infectious diseases [57]. The CovidQA dataset is for questions and answers about COVID-19 and is

built from the source from Kaggle's COVID-19 open research dataset challenge. The CovidQA dataset is the first publicly available COVID-19 Q&A dataset, which contains 124 question-article pairs according to the CovidQA dataset. The current version is 0.1, and the database will be further expanded as resources get evaluated. The dataset includes fields such as subcategory, title, answer. Title is the title of the scientific research article, of which the answer is derived or the title of the announcement issued by an authoritative institution to verify the reliability of the answer. The datasets were reviewed by epidemiologists, MDs (Medical Practitioners), and medical students. The dataset can be used in the natural language processing model (NLP) field to test the validity of the model. In addition, the dataset can be used to build the deep learning-based question and answer model, which consists of two main parts: the question context component and the answer component [29].

University College London recently published a dataset called RWWD (Real World Worry Dataset) [25]. The dataset utilizes a direct questionnaire approach to obtain written descriptions of how people feel about COVID-19 and their current emotions. Instead of relying on a third party to annotate, the dataset relies on the writer's ratings of their own mood after writing, which makes the dataset more reliable. RWWD has two versions, and each version has 2,500 English texts. The first version is of variable length, with a minimum of 500 words. The second version uses the Twitter format (i.e., no more than 240 words long), and the short text is mainly used for comparison with the Twitter data. All subjects were asked to use a 9-point scales to indicate their internal emotions including worry, anger, etc. The study results on this dataset showed that Britons were more worried about their families and finances. Short texts (in the form of tweets) tend to be inspirational and chants, while long texts prefer to express their inner emotions, for example, people's concerns about the epidemic. This dataset has been used to measure changes in the mood of citizens during the COVID-19 outbreak.

5. Discussions

More and more datasets related to COVID-19 pandemic are emerging gradually over time. However, only a part of datasets are helpful for researchers as most the published datasets for COVID-19 analysis or treatments tend to be incomplete, possibly biased, and limited to national samples. Especially, problems existing in COVID-19 datasets such as incompleteness [23] and small scale [29] are urgent problem for research. Thus, how to obtain valuable datasets is still a challenging task for researchers. For the sake of effective COVID-19 research, we still need more valuable datasets by adopting appropriate processing and collection methods.

Nevertheless, the presented datasets have significant implications to fight against COVID-19. For instance, real-time datasets can be used for contact tracing and finding out the influence scope. These datasets can help publish reasonable policy of lockdown. Similar, these datasets can be used to identify potential risk points such as, places for common public interests. The listed datasets can have practical implications when integrated with the other datasets. For instance, real-time datasets can provide information regarding the impact of diseases on human life. These datasets can be used with other datasets including temperature and humidity. These factors appear to influence the COVID-19 effects on

human lives [49]. Future studies can use these factors to reveal their moderating effect between COVID-19 and human society regarding deaths and population movements.

Companies can use knowledge bases to develop strategies to fight against diseases. The listed knowledge bases are a valuable tool to extract practical knowledge, experience and facts to formulate policies for the businesses. For instance, companies can use these datasets to publish economic policy recommendations as small and medium businesses are heavily impacted by diseases. The business survival depends on the new business models and how quickly these models are adopted. Therefore, new strategies are inevitable for governments and local businesses to endure this unpredicted scenario [7].

Researchers can use the media datasets to publish policies to counter fake news regarding the COVID-19. Online platforms are flooded with unauthentic misinformation such as the negative impacts of the COVID-19 vaccine and the dangerous nature of the virus [32]. This fabricated news can make negative impact on the efforts to counter the disease. Firstly, people might not take the required precautions and consider the COVID-19 a conspiracy [3], as witnessed in several anti-lockdown protests. Second, people might resist receiving vaccinations based on social media fake news. This is also evident with the slow progress of vaccination in several countries.

Social media plays a critical role in addressing the issue of misinformation. It can also be a valuable tool to provide relevant and authentic information for patients, doctors and clients. Therefore, better social media approaches and strategies are required for social media to play an influential role in policy and decision making for government, organizations and individuals [11]. The listed media datasets can provide researchers with a platform to generate recommendations of such policies and strategies.

6. Conclusion

This paper presents several key sources of COVID-19 datasets of different categories including time-series datasets (real-time cases datasets, migration information datasets), knowledge base (knowledge graphs, research dataset), and media datasets (social networks, X-ray image and CT scan datasets, news, and others). Then, we discusses how various organisations gather the data. Most of the datasets examined in this paper are publicly available. We provide relevant links to the datasets wherever possible. We also have discussed several drawbacks of the current COVID-19 datasets. An efficient COVID-19 dataset evaluation procedure is also missing. We suggest building a more effective mechanism for collecting more valuable data for research.

Acknowledgments. The authors would like to thank Xiangtai Chen, Huazhu Cao, Mengyuan Wang, and Xu Feng from Dalian University of Technology for their help with the first draft of this paper.

References

1. Al-Saleem, J., Granet, R., Ramakrishnan, S., Ciancetta, N.A., Saveson, C., Gessner, C., Zhou, Q.: Knowledge graph-based approaches to drug repurposing for covid-19. Journal of Chemical Information and Modeling 61(8), 4058–4067 (2021)

2. Albahri, A., Hamid, R.A., Alwan, J.K., Al-Qays, Z., Zaidan, A., Zaidan, B., Albahri, A., AlAmoodi, A., Khlaf, J.M., Almahdi, E., et al.: Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (covid-19): a systematic review. Journal of Medical Systems 44, 1–11 (2020)
3. Apuke, O.D., Omar, B.: Fake news and covid-19: modelling the predictors of fake news sharing among social media users. Telematics and Informatics 56, 101475 (2021)
4. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., Chowell, G.: A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. Epidemiologia 2(3), 315–324 (2021), https://www.mdpi.com/2673-3986/2/3/24
5. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., Ciccozzi, M.: Application of the arima model on the covid-2019 epidemic dataset. Data in Brief p. 105340 (2020)
6. Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., Zheng, J.: The psychological impact of the covid-19 epidemic on college students in china. Psychiatry Research p. 112934 (2020)
7. Carracedo, P., Puertas, R., Marti, L.: Research lines on the impact of the covid-19 pandemic on business. a text mining analysis. Journal of Business Research 132, 586–593 (2021)
8. Chen, E., Lerman, K., Ferrara, E., et al.: Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance 6(2), e19273 (2020)
9. Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A.P., Mu, K., Rossi, L., Sun, K., et al.: The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. Science 368(6489), 395–400 (2020)
10. Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M.: Covid-19 image data collection: Prospective predictions are the future. arXiv 2006.11988 (2020), https://github.com/ieee8023/covid-chestxray-dataset
11. Cuello-Garcia, C., Pérez-Gaxiola, G., van Amelsvoort, L.: Social media can have an impact on how we manage and investigate the covid-19 pandemic. Journal of Clinical Epidemiology 127, 198–201 (2020)
12. Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., Kodamullil, A.T.: Covid-19 knowledge graph: a computable, multimodal, cause-and-effect knowledge model of covid-19 pathophysiology. Bioinformatics 37(9), 1332–1334 (2021)
13. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track covid-19 in real time. The Lancet Infectious Diseases 20(5), 533–534 (2020)
14. Ferreira, C.M., Sá, M.J., Martins, J.G., Serpa, S.: The covid-19 contagion–pandemic dyad: A view from social sciences. Societies 10(4), 77 (2020)
15. Gao, Z., Yada, S., Wakamiya, S., Aramaki, E.: Naist covid: Multilingual covid-19 twitter and weibo dataset. arXiv preprint arXiv:2004.08145 (2020)
16. Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., Yang, H., Hong, L., Wu, N., Yuan, E., et al.: A data-driven drug repositioning framework discovered a potential therapeutic agent targeting covid-19. BioRxiv (2020)
17. Hamzah, F.B., Lau, C., Nazri, H., Ligot, D.V., Lee, G., Tan, C.L., Shaib, M., Zaidon, U.H.B., Abdullah, A.B., Chung, M.H., et al.: Coronatracker: worldwide covid-19 outbreak data analysis and prediction. Bull World Health Organ 1(32), 1–32 (2020)
18. Hong, L., Lin, J., Tao, J., Zeng, J.: Bere: An accurate distantly supervised biomedical entity relation extraction network. arXiv preprint arXiv:1906.06916 (2019)
19. Hou, M., Ren, J., Zhang, D., Kong, X., Zhang, D., Xia, F.: Network embedding: Taxonomies, frameworks and applications. Computer Science Review 38, 100296 (2020)
20. Huang, J., Wang, H., Xiong, H., Fan, M., Zhuo, A., Li, Y., Dou, D.: Quantifying the economic impact of covid-19 in mainland china using human mobility data. arXiv preprint arXiv:2005.03010 (2020)

21. Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., Kaur, M.: Classification of the covid-19 infected patients using densenet201 based deep transfer learning. Journal of Biomolecular Structure and Dynamics pp. 1–8 (2020)
22. Kabir, M., Madria, S., et al.: Coronavis: A real-time covid-19 tweets analyzer. arXiv preprint arXiv:2004.13932 (2020)
23. Karlinsky, A., Kobak, D.: Tracking excess mortality across countries during the covid-19 pandemic with the world mortality dataset. Elife 10, e69336 (2021)
24. Karmakar, M., Lantz, P.M., Tipirneni, R.: Association of social and demographic factors with covid-19 incidence and death rates in the us. JAMA Network Open 4(1), e2036462–e2036462 (2021)
25. Kleinberg, B., van der Vegt, I., Mozes, M.: Measuring emotions in the covid-19 real world worry dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (2020)
26. Lai, S., Bogoch, I.I., Ruktanonchai, N.W., Watts, A., Lu, X., Yang, W., Yu, H., Khan, K., Tatem, A.J.: Assessing spread risk of wuhan novel coronavirus within and beyond china, january-april 2020: a travel network-based modelling study. MedRxiv (2020)
27. LaiS, H., et al.: Changingepidemiologyofhug man brucellosis, china, 1955g2014. EmergInfect Dis 23(2), 184 (2017)
28. Lee, J., Sean, S.Y., Jeong, M., Sung, M., Yoon, W., Choi, Y., Ko, M., Kang, J.: Answering questions on covid-19 in real-time. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020 (2020)
29. Levy, S., Mo, K., Xiong, W., Wang, W.Y.: Open-domain question-answering for covid-19 and other emergent domains. arXiv preprint arXiv:2110.06962 (2021)
30. Li, X., Geng, M., Peng, Y., Meng, L., Lu, S.: Molecular immune pathogenesis and diagnosis of covid-19. Journal of Pharmaceutical Analysis (2020)
31. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Synthesis Lectures on Human Language Technologies 14(4), 1–325 (2021)
32. van der Linden, S., Roozenbeek, J., Compton, J.: Inoculating against fake news about covid-19. Frontiers in Psychology 11, 2928 (2020)
33. Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., Lee, I.: Artificial intelligence in the 21st century. IEEE Access 6, 34403–34421 (2018)
34. Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., Xu, B., Xia, F.: Data mining and information retrieval in the 21st century: A bibliographic review. Computer Science Review 34, 100193 (2019)
35. Liu, J., Nie, H., Li, S., Chen, X., Cao, H., Ren, J., Lee, I., Xia, F.: Tracing the pace of covid-19 research: Topic modeling and evolution. Big Data Research 25, 100236 (2021)
36. Liu, J., Ren, J., Zheng, W., Chi, L., Lee, I., Xia, F.: Web of scholars: A scholar knowledge graph. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2153–2156 (2020)
37. Liu, J., Tian, J., Kong, X., Lee, I., Xia, F.: Two decades of information systems: a bibliometric review. Scientometrics 118(2), 617–643 (2019)
38. Mandal, M., Jana, S., Nandi, S.K., Khatua, A., Adak, S., Kar, T.: A model based study on the dynamics of covid-19: Prediction and control. Chaos, Solitons & Fractals p. 109889 (2020)
39. Mohamadou, Y., Halidou, A., Kapen, P.T.: A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19. Applied Intelligence pp. 1–13 (2020)
40. Möller, T., Reina, A., Jayakumar, R., Pietsch, M.: Covid-qa: A question answering dataset for covid-19. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (2020)
41. Naseem, U., Razzak, I., Khushi, M., Eklund, P.W., Kim, J.: Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. IEEE Transactions on Computational Social Systems (2021)
42. Nemes, L., Kiss, A.: Social media sentiment analysis based on covid-19. Journal of Information and Telecommunication pp. 1–15 (2020)

43. Ngai, H., Park, Y., Chen, J., Parsapoor, M.: Transformer-based models for question answering on covid19. arXiv preprint arXiv:2101.11432 (2021)
44. Niehus, R., De Salazar, P.M., Taylor, A.R., Lipsitch, M.: Using observational data to quantify bias of traveller-derived covid-19 prevalence estimates in wuhan, china. The Lancet Infectious Diseases (2020)
45. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of covid-19 cases using deep neural networks with x-ray images. Computers in Biology and Medicine 121, 103792 (2020)
46. Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C., Tizzoni, M.: Covid-19 outbreak response, a dataset to assess mobility changes in italy following national lockdown. Scientific Data 7(1), 1–7 (2020)
47. Prakash, K.B., Imambi, S.S., Ismail, M., Kumar, T.P., Pawan, Y.: Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms. International Journal 8(5) (2020)
48. Punn, N.S., Sonbhadra, S.K., Agarwal, S.: Covid-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv (2020)
49. Qi, H., Xiao, S., Shi, R., Ward, M.P., Chen, Y., Tu, W., Su, Q., Wang, W., Wang, X., Zhang, Z.: Covid-19 transmission in mainland china is associated with temperature and humidity: A time-series analysis. Science of The Total Environment 728, 138778 (2020)
50. Ren, J., Xia, F., Chen, X., Liu, J., Hou, M., Shehzad, A., Sultanova, N., Kong, X.: Matching algorithms: Fundamentals, applications and challenges. IEEE Transactions on Emerging Topics in Computational Intelligence 5(3), 332–350 (2021)
51. Roda, W.C., Varughese, M.B., Han, D., Li, M.Y.: Why is it difficult to accurately predict the covid-19 epidemic? Infectious Disease Modelling (2020)
52. Shahi, G.K., Nandini, D.: Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. arXiv preprint arXiv:2006.11343 (2020)
53. Shen, I., Zhang, L., Lian, J., Wu, C.H., Fierro, M.G., Argyriou, A., Wu, T.: In search for a cure: recommendation with knowledge graph on cord-19. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3519–3520 (2020)
54. Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, R.: World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). International Journal of Surgery (2020)
55. Su, D., Xu, Y., Yu, T., Siddique, F.B., Barezi, E.J., Fung, P.: Caire-covid: a question answering and query-focused multi-document summarization system for covid-19 scholarly information management. arXiv preprint arXiv:2005.03975 (2020)
56. Sun, K., Wang, L., Xu, B., Zhao, W., Teng, S.W., Xia, F.: Network representation learning: From traditional feature learning to deep learning. IEEE Access 8, 205600–205617 (2020)
57. Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K., Lin, J.: Rapidly bootstrapping a question answering dataset for covid-19. arXiv preprint arXiv:2004.11339 (2020)
58. Tátrai, D., Várallyay, Z.: Covid-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability. arXiv preprint arXiv:2003.14160 (2020)
59. Tuli, S., Tuli, S., Tuli, R., Gill, S.S.: Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing. Internet of Things 11, 100222 (2020)
60. Ucar, F., Korkmaz, D.: Covidiagnosis-net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from x-ray images. Medical Hypotheses 140, 109761 (2020)
61. Ullah, A., Das, A., Das, A., Kabir, M.A., Shu, K.: A survey of covid-19 misinformation: Datasets, detection techniques and open issues. arXiv preprint arXiv:2110.00737 (2021)
62. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports 10(1), 1–12 (2020)

63. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R.M., et al.: Cord-19: The covid-19 open research dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (2020)
64. Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., et al.: A deep learning algorithm using ct images to screen for corona virus disease (covid-19). European Radiology pp. 1–9 (2021), https://doi.org/10.1016/j.mehy.2020.109761
65. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
66. Wang, X., Song, X., Li, B., Guan, Y., Han, J.: Comprehensive named entity recognition on cord-19 with distant or weak supervision. arXiv preprint arXiv:2003.12218 (2020)
67. Warren, M.S., Skillman, S.W.: Mobility changes in response to covid-19. arXiv preprint arXiv:2003.14228 (2020)
68. Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Bonten, M.M., Dahly, D.L., Damen, J.A., Debray, T.P., et al.: Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. British Medical Journal 369 (2020)
69. Xia, F., Jedari, B., Yang, L.T., Ma, J., Huang, R.: A signaling game for uncertain data delivery in selfish mobile social networks. IEEE Transactions on Computational Social Systems 3(2), 100–112 (2016)
70. Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., Liu, H.: Graph learning: A survey. IEEE Transactions on Artificial Intelligence 2(2), 109–127 (2021)
71. Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., Liu, C.: Exploring human mobility patterns in urban scenarios: A trajectory data perspective. IEEE Communications Magazine 56(3), 142–149 (2018)
72. Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., Cohn, E.L., Hswen, Y., Hill, S.C., Cobo, M.M., et al.: Epidemiological data from the covid-19 outbreak, real-time case information. Scientific Data 7(1), 1–6 (2020)
73. Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J.F., Li, X., Xu, W., Torvik, V.I., et al.: Building a pubmed knowledge graph. Scientific Data 7(1), 1–15 (2020)
74. Yang, C., Zhou, X., Zafarani, R.: Checked: Chinese covid-19 fake news dataset. Social Network Analysis and Mining 11(1), 1–8 (2021)
75. Yu, J.: Open access institutional and news media tweet dataset for covid-19 social science research. arXiv preprint arXiv:2004.01791 (2020)
76. Yu, S., Qing, Q., Zhang, C., Shehzad, A., Oatley, G., Xia, F.: Data-driven decision-making in covid-19 response: A survey. IEEE Transactions on Computational Social Systems 8(4), 1016–1029 (2021)
77. Zarei, K., Farahbakhsh, R., Crespi, N., Tyson, G.: A first instagram dataset on covid-19. arXiv preprint arXiv:2004.12226 (2020)
78. Zeroual, A., Harrou, F., Dairi, A., Sun, Y.: Deep learning methods for forecasting covid-19 time-series data: A comparative study. Chaos, Solitons & Fractals 140, 110121 (2020)
79. Zhang, D., Zhang, M., Peng, C., Jung, J.J., Xia, F.: Metaphor research in the 21st century: A bibliographic analysis. Computer Science and Information Systems 18, 303–322 (2021)
80. Zhang, J., Wang, W., Xia, F., Lin, Y.R., Tong, H.: Data-driven computational social science: A survey. Big Data Research p. 100145 (2020)
81. Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., Kilicoglu, H.: Drug repurposing for covid-19 via knowledge graph completion. Journal of Biomedical Informatics 115, 103696 (2021)
82. Zhao, J., Zhang, Y., He, X., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020)

83. Zong, S., Baheti, A., Xu, W., Ritter, A.: Extracting covid-19 events from twitter. arXiv preprint arXiv:2006.02567 (2020)

Ke Sun received the B.Sc. and M.Sc. degrees from Shandong Normal University, Jinan, China. He is currently Ph.D. Candidate in Software Engineering at Dalian University of Technology, Dalian, China. His research interests include deep learning, network representation learning, and knowledge graphs.

Wuyang Li is currently working toward the Bachelor's degree in the School of Software, Dalian University of Technology, China. His research interests include data science, natural language processing, and social network analysis.

Vidya Saikrishna received her PhD from Monash University, Australia in 2017. She completed her Master's in Technology (M. Tech) in 2012 and Bachelor's in Engineering in 2002 from Maulana Azad National Institute of Technology, India and Barkatullah University, India respectively. She is currently a Scholarly Teaching Fellow in Global Professional School, Federation University Australia. Her current research interests include Machine Learning, Artificial Intelligence, String Matching, and Data/Text Mining.

Mehmood Chadhar received his PhD in Information Systems from the University of New South Wales, Sydney, Australia. He is currently a Lecturer teaching business analytics, supply chain management, and real-time analytics at Federation University Australia. His areas of interest include enterprise systems implementation, organizational learning, IT business value and social media benefits.

Feng Xia received the BSc and PhD degrees from Zhejiang University, Hangzhou, China. He was Full Professor and Associate Dean (Research) in School of Software, Dalian University of Technology, China. He is Associate Professor and former Discipline Leader (IT) in Institute of Innovation, Science and Sustainability, Federation University Australia. Dr. Xia has published 2 books and over 300 scientific papers in international journals and conferences (such as IEEE TAI, TKDE, TNNLS, TBD, TCSS, TNSE, TETCI, TC, TMC, TPDS, TETC, THMS, TVT, TITS, TASE, ACM TKDD, TIST, TWEB, TOMM, WWW, AAAI, SIGIR, CIKM, JCDL, EMNLP, and INFOCOM). His research interests include data science, artificial intelligence, graph learning, anomaly detection, and systems engineering. He is a Senior Member of IEEE and ACM.

Received: August 22, 2021; Accepted: April 22, 2022.