Big Data Research 25 (2021) 100236

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Tracing the Pace of COVID-19 Research: Topic Modeling and Evolution

Jiaying Liu [a], Hansong Nie [a], Shihao Li [a], Xiangtai Chen [a], Huazhu Cao [a], Jing Ren [b], Ivan Lee [c], Feng Xia [b,*]

[a] School of Software, Dalian University of Technology, Dalian 116620, China
[b] School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia
[c] STEM, University of South Australia, Adelaide 5001, Australia

ARTICLE INFO

Article history:
Received 16 September 2020
Received in revised form 12 March 2021
Accepted 13 April 2021
Available online 24 April 2021

Keywords:
COVID-19
Deep learning
Topic modeling
Bibliometric analysis
Science of Science

ABSTRACT

COVID-19 has been spreading rapidly around the world. With the growing attention on the deadly pandemic, discussions and research on COVID-19 are rapidly increasing to exchange latest findings with the hope to accelerate the pace of finding a cure. As a branch of information technology, artificial intelligence (AI) has greatly expedited the development of human society. In this paper, we investigate and visualize the on-going advancements of early scientific research on COVID-19 from the perspective of AI. By adopting the Latent Dirichlet Allocation (LDA) model, this paper allocates the research articles into 50 key research topics pertinent to COVID-19 according to their abstracts. We present an overview of early studies of the COVID-19 crisis at different scales including referencing/citation behavior, topic variation and their inner interactions. We also identify innovative papers that are regarded as the cornerstones in the development of COVID-19 research. The results unveil the focus of scientific research, thereby giving deep insights into how the academic society contributes to combating the COVID-19 pandemic.

1. Introduction

COVID-19, the pneumonia caused by 2019 new coronavirus infection, has been declared as a pandemic by the World Health Organization (WHO), with over 110 million confirmed cases all over the world as of early March 2021. The continuous growth of the COVID-19 pandemic has attracted attention from researchers worldwide. During this pandemic with high infection rate, research communities, funding agencies, and practitioners are working together to better control, mitigate, and suppress its spread. Most scholars have been committed to the study of COVID-19 transmission [1–3], outbreak detection and analysis [4,5], vaccines and treatments [6], and socio-economic impact analysis [7]. In addition, public, private and non-profit organizations have taken different initiatives to publicly share COVID-19 related scientific research. WHO and European Commission provide relevant information to the public every day. Similarly, COVID-19 Open Research Dataset (CORD-19)[1] is a weekly-updated dataset for publications and research achievements on both COVID-19 and other coronaviruses.

It is important to review the current research situation during the on-going COVID-19 pandemic. On the one hand, it could help researchers understand how existing techniques such as machine learning and artificial intelligence (AI) could help the global response to the COVID-19 pandemic. On the other hand, related theoretical reference give insights into the decision-making of government, such as “social distancing”, “school closure”, and “telecommuting”. The analysis of scientific publications, especially theme or topic analysis, is usually considered as the key proxy of reflecting the trend of research in both theory and practice. Until now, a lot of scientific research have been devoted to studying publication data from the perspectives of topic modeling and analysis. For example, Sun et al. [8] reveal the evolution of transportation research relying on the ability of topic modeling techniques. Liu et al. [9] profile the field of Information Systems and highlight its evolution by studying its main trends from 1996 to 2015 at the age of Big Scholarly Data [10]. Topic modeling is also applied in COVID-19 research to help reflect people’s emotional response to the coronavirus and academic research hotspots. Jang et al. [11] analyze COVID-19 related tweets using topic modeling and aspect-based sentiment analysis to investigate people’s reactions and concerns about COVID-19 in North America and Canada. Similarly, Ordun et al. [12] also trace the distinctiveness of topics, key terms and features, speed of information dissemination, and network behaviors for COVID-19 tweets by using Latent Dirichlet Allocation (LDA) to

* Corresponding author.
E-mail address: f.xia@ieee.org (F. Xia).

[1] https://www.semanticscholar.org/cord19.

https://doi.org/10.1016/j.bdr.2021.100236
2214-5796/

generate topics. From the perspective of scientific article analysis, Sonbhadra et al. [13] extract the activity and trends of coronavirus related research articles using machine learning approaches from the paper content level. Dong et al. [14] use LDA to track semantic relationships between topics and compare the topic distribution between COVID-19 and other CoV infections. Different from these related studies, we not only identify the topics of early COVID-19 studies, but also track the temporal changes and internal relationship of these topics. More importantly, we also propose relevant methods to predict innovative topics, which have not been illustrated in other studies.

In order to study how early scientific research reacts to the pandemic, this paper present a scientometric analysis of COVID-19 related research based on artificial intelligence technology. The study can not only show important issues that scholars currently pay attention to, but also assist in predicting future research directions. We have conducted an extensive review of emerging literature to better understand the change in research context over time. To be more specific, we profile COVID-19 related research to explore answers of the following questions:

- What are the main areas of COVID-19 related publications? How do these fields change over time? Which areas of research are related to COVID-19 research?
- Which topics are of great concern to the authors in COVID-19 research? How do they relate to the on-going COVID-19 pandemic? How do these topics interact with each other?
- What problems for COVID-19 research have not been solved so far? What are potential future directions for the on-going COVID-19 pandemic?

The remainder of this study is organized as follows. Section 2 provides details of methodology used in this study, including data collection and pre-processing, as well as topic discovery techniques. We also list several metrics to measure the evolution of topics. Section 3 summarizes findings for extensive analysis of topic distribution and evolution over time. Finally, we conclude the study and list some limitations and potential future directions in Section 4.

2. Methodology

To give deep insights into COVID-19 related research, we obtain a large-scale scholarly dataset from the COVID-19 Open Research Dataset (CORD-19) [15]. CORD-19 is a growing resource containing not only scientific papers on COVID-19, but also relevant research achievement on other types of coronavirus. The publication metadata are collected from PubMed Central (PMC),[2] BioRXiv preprint servers, MedRxiv preprint servers, and WHO COVID-19 Database. Due to the scope of this review being restricted to COVID-19 related publications, the selection of the CORD-19 papers for analysis are limited to the papers published after 2020, with the title or abstract containing keywords “COVID-19”, “SARS-CoV-2”, or “2019-NCOV”. In order to classify papers by research fields, we match “Microsoft Academic Paper ID” of CORD-19 dataset in Microsoft Academic Graph provided by Microsoft Cognitive Services [16]. Finally, our corpus of study is comprised of a total of 6392 publications, including 714 preprinted papers from arXiv, 1307 papers from PubMed, 1648 papers from Elsevier, 66 papers from CHI, 480 papers from WHO, 2162 papers from MedRXiv, and 555 papers from BioRXiv.

2.1. Topic modeling and evolution

2.1.1. Topic modeling

Latent Dirichlet Allocation (LDA) is a generative probability model based on Bayesian learning proposed by Blei et al. [17]. It is used to generate topics from documents and has been widely used in text data mining and biological information processing. To quantify the variation of topics across COVID-19 research, we use LDA for topic modeling.

The LDA model consists of a word set W, a document set D, and a topic set Z. In the word set $W = \{w_1, ..., w_v, ..., w_V\}$, v represents the index of words, V is the number of words. For the document set $D = \{d_1, ..., d_m, ..., d_M\}$, m is the index of documents and M represents the number of documents. The document d_m is a word sequence $d_m = (d_{m1}, ..., d_{mn}, ..., d_{mN_m})$, where d_{mn} is the n-*th* word in d_m. N_m is the number of words in d_m. For the topic set $Z = \{z_1, ..., z_k, ..., z_K\}$, k is the index of topics and K is the number of topics.

Each topic z_k is determined by the conditional distribution probability $p(w|z_k)$ of a word $w \in W$. The distribution $p(w|z_k)$ obeys the multinomial distribution with the parameter φ_k. φ_k is picked from a Dirichlet distribution with a hyper-parameter β. $\varphi_k = (\varphi_{k1}, \varphi_{k2}, ..., \varphi_{kV})$ is a V-dimensional vector, where φ_{kv} is the probability of topic z_k generating word w_v. Each document d_m is determined by the conditional probability distribution $p(z|d_m)$ of a topic z, $w \in Z$. The distribution $p(z|d_m)$ obeys the multinomial distribution with the parameter θ_m. θ_m is picked from a Dirichlet distribution with a hyper-parameter α. $\theta_m = (\theta_{m1}, \theta_{m2}, ..., \theta_{mK})$ is a K-dimensional vector, where θ_{mk} is the probability of document d_m generating topic z_k. Each word w_{mn} in each document d_m is determined by the topic distribution $p(z|d_m)$ of the document and the word distribution $p(w|z_k)$ of all topics.

The generation process of LDA model is described as follows.

- **Generate K topics**. A φ_k is picked from a Dirichlet distribution $Dir(\beta)$ to generate the word distribution $p(w|z_k)$ of topic z_k over W.
- **Generate topic distribution**. A θ_m is picked from a Dirichlet distribution $Dir(\alpha)$ to generate the topic distribution $p(z|d_m)$ of document d_m.
- **Generate word sequences**. It firstly samples particular topic z_{mn} from a multinomial distribution $Mult(\theta_m)$, and then the word w_{mn} is selected from a multinomial distribution $Mult(\varphi_{z_{mn}})$.

LDA model takes the abstract of each paper as input, and the data pre-processing steps are listed as follows:

1. Use Langid[3] to identify English papers.
2. Convert all uppercase initials in abstracts to lowercase.
3. Divide sentences into independent words.
4. Remove stopwords such as “is, the, have”.
5. Convert documents into bags-of-words representations.

After data pre-processing, LDA infers the posterior of document-topic distribution θ and topic-word distribution φ effectively. Finally, we generate 50 topics in total and infer the topic distribution of each paper.

2.1.2. Topic evolution metrics

After inferring the topic names, we first analyze the cumulative proportion of topics. For each topic z_k, the cumulative proportion c_k is calculated as follows:

[2] https://www.ncbi.nlm.nih.gov/pmc/.

[3] https://github.com/saffsd/langid.py.

$$c_k = \sum_{m=1}^{M} \theta_{mk} \tag{1}$$

where θ_{mk} is the proportion of topic z_k in paper d_m.

In order to analyze the topic distribution over time, 7-day windows are applied in between January 1st and May 20th. We analyze the time variation of topics by calculating the proportion of each topic in each time period. The proportion $c_k^{[t]}$ of topic z_k in time period p_t is calculated as follows:

$$c_k^{[t]} = \sum_{m=1}^{M} \theta_{mk} \times \mathbb{I}(T_m \in p_t) \tag{2}$$

where $\mathbb{I}(e) = 1$ if e is true, otherwise, $\mathbb{I}(e) = 0$. T_m is the submission date of paper d_m.

To identify topics with growing popularity, we compute the popularity r_k of topics every month from February to May. The popularity of each topic in the current month is the ratio of its proportion in the current month to that in the previous month. For example, the popularity r_k of topic z_k in February is computed as follows:

$$r_k^{(Feb)} = c_k^{[Feb]} / c_k^{[Jan]} \tag{3}$$

where $c_k^{[Feb]}$ is the topic proportion in February, $c_k^{[Jan]}$ is the topic proportion in January. The sum of the monthly popularity of each topic is the total popularity r_k of the topic. The formula of r_k is as follows:

$$r_k = r_k^{(Feb)} + r_k^{(Mar)} + r_k^{(Apr)} + r_k^{(May)}. \tag{4}$$

$r_k^m > 1$ indicates that the proportion of topic z_k has increased from the month $m-1$ to the month m. These topics can be regarded as hot topics. On the contrary, $r_k^m < 1$ indicates that the proportion of topic z_k has decreased. We call these topics as unpopular topics.

2.2. Innovative paper identification

In order to identify milestones and key changes in COVID-19 research, we want to find innovative topics and papers. We first label papers according to their submission time. The label of papers submitted in January 2020 is 0, because only a few papers are submitted in this month. The label is incremented by 1 every seven days since February 1st. For example, the label of papers submitted on February 1st to February 7th is 1, and papers submitted from February 8th to February 14th are labeled as 2. The maximal label is 14. Following the steps mentioned in Section 2.1.1, we get representations of papers by 50-dimensional vectors reflecting their topic distribution. These vectors reflect the similarity between papers from topic level. We use v_p and L_p to denote the vector representation and label of paper p, respectively.

We then use the K-Means clustering algorithm to divide papers into 15 classes. K-Means is an unsupervised method and performs well on unbalanced data, where papers of one class may outnumber papers of another class. We give each class c a label L_c according to labels of corresponding papers:

$$L_c = argmax_y |\{p | p \in \mathcal{P}_y \wedge p \in \mathcal{P}_c\}| \tag{5}$$

where y is the label of papers. $\mathcal{P}_y$ and $\mathcal{P}_c$ are sets of papers of label y and class c, respectively. As shown in the above equation, L_c is the label with the most papers of class c. We define I_p as the time interval of paper p and papers that are similar to it:

$$I_p = L_{c_p} - L_p \tag{6}$$

where c_p is the class of paper p predicted by K-Means. In this paper, two papers are considered similar if they fall under the same class, i.e., similar topic distribution. I_p reflects foreseeing ability of paper p. A greater I_p means greater foreseeing ability. A paper is considered innovative if its topic distribution matches the topic distribution of papers published in the future, especially in the distant future [18]. Therefore, I_p also reflects the innovative ability of paper p.

An innovative paper is considered the cornerstone with seminal work followed by many papers, and thus its topic vector is a center of vectors of the corresponding papers. We define the topic vector of one class as the average vector of its papers:

$$v_c = \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} v_p. \tag{7}$$

The distance between v_p and v_{c_p} is defined as:

$$D_p = ||v_p - v_{c_p}||_1 \tag{8}$$

where $||v||_1$ is the $L1$-norm of vector v. Smaller D_p means greater probability of being the center of class c_p.

Last, we propose an index named IA to measure innovative ability. Specifically, the IA of paper p is defined as:

$$IA_p = J_p * I_p / D_p \tag{9}$$

where J_p is the normalized journal impact factor for paper p published on journal J. If the paper is a preprint, J_p is set as the average of all normalized journal impact factors. The index considers both the publication year, journal importance, and the position among similar papers. In this paper, the top ten papers that have maximal IA are considered as innovative papers.

We identify innovative topics through innovative papers, that is, topics from innovative papers are also innovative. LDA represents a passage by topic distributions and represents topics by word distributions. Topic distributions are latent variables and cannot be characterized by a simple word. For simplicity, we use some keywords to describe the topics of one paper. In this paper, the topics of one paper are represented by ten words that have the most probability to appear in the paper. For paper p, the probability of word w to appear is calculated as:

$$pro_{pw} = \sum_t \sum_w pro_{pt} pro_{tw} \tag{10}$$

where pro_{pt} is the probability that paper p is generated via topic t. pro_{tw} is the probability that word w appears in topic t. These two variables are obtained from LDA.

3. Results and analysis

This section focuses on profiling COVID-19 research using defined measures and explores the following aspects: (1) Analyze the research field of COVID-19 publications and their citation behavior. (2) Detect topics variation over time and explore their inner connections. (3) Identify innovative papers and topics.

3.1. Fields of research

Fig. 1 presents the weekly number of publications during the period of study. As shown in Fig. 1, only few papers have been published during the first several weeks. Then the number of papers has grown dramatically since mid-March 2020. Nearly 700 papers have been published every week in April. We divide all papers into 19 fields. Fig. 2 shows the number of publications in

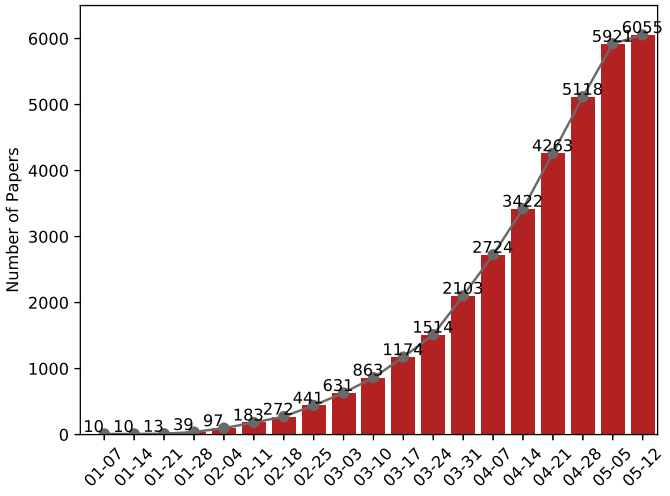


Fig. 1. Cumulative number of publications per week.

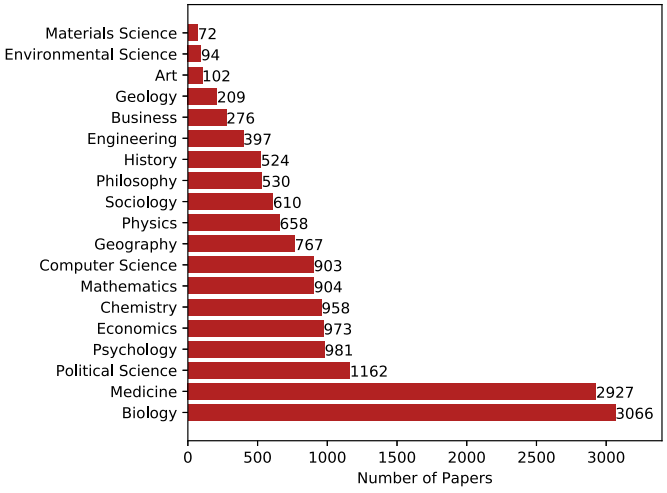


Fig. 2. Number of publications in each field.

different research fields. From the results we can see that most papers belong to Biology and Medicine. They accounted for 18% (Biology) and 19% (Medicine) of all publications, respectively. Beyond that, scholars pay more attention to psychological and economic research. In addition to coronavirus mechanisms and drug discovery, the most common applications of the scientific response to COVID-19 in the area include sentiment analysis, social and economic impact analysis, and outbreak detection. Overall, research contributions of COVID-19 span across diverse aspects of scientific fields.

To capture the temporal dynamics of papers in different fields and measure the overall development of COVID-19 research during the period of study, we uncover the correlations between topics and time. As shown in Table 1, Biology and Medicine papers account for the largest proportion since the first week. The proportion of Psychology and Political Science papers shows a rapid growth trend, especially in the second week of May. It suggests that at the early stage of the study, scholars pay more attention to Biology and Medicine related research such as protein structure prediction, drug repurposing, and gene expression signatures. With the passage of time, the analysis of social, economic, and psychological impacts caused by the coronavirus also attracts a lot of attention. In summary, with the increase of diversity in research fields, the proportion of publications in Medicine and Biology are decreasing, and the number of psychological research is steadily

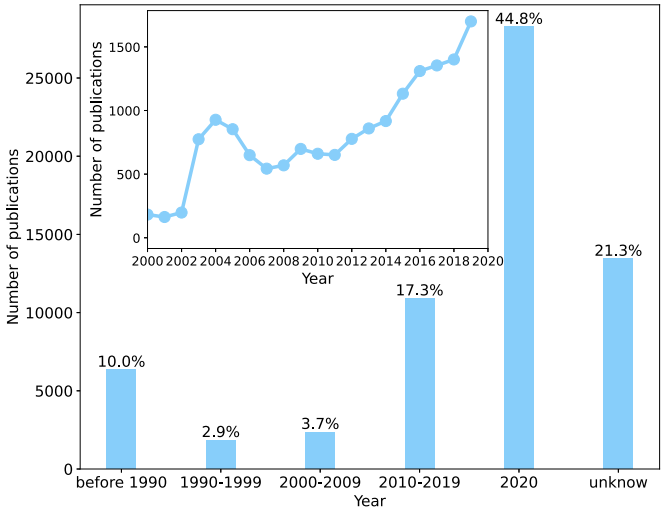


Fig. 3. Published year distribution of references in COVID-19 related research.

increasing. Public health emergencies have also sparked fears of an impending economic crisis and recession. As the literature illustrated [19,20], lockdown measures such as social distancing, self-isolation, and travel restrictions have not only lead to a reduced workforce across all economic sectors but also affected the mental health of individuals from the perspective of insecurity, confusion, and emotional isolation. Consequently, research on mental health and economic impact has also greatly increased.

3.2. Citation behavior analysis

We analyze the citations (i.e., references) of COVID-19 research from two aspects: published year distribution, and fields of research. After screening, the corpus of COVID-19 papers has more than 118,000 references, including more than 60,000 non-repetitive articles. Fig. 3 shows the published year distribution of all references and the proportions of references that belong to each decade. We can also find that the number of references published in 2002 has grown dramatically. The SARS pandemic in 2002 has drawn great attention to the research on coronavirus. The main reference object of COVID-19 is the SARS virus because of close similarity, and COVID-19 papers mostly cites SARS papers that are published after 2002. Among these papers, most literature published in 2010-2019, twice that of the previous decade. This is because during the outbreak of the Middle East Respiratory Syndrome (MERS) in 2012, interests in coronaviruses boomed again.

From the perspective of research fields, these references cover more than 30,000 subfields such as virology, medicine, biology, immunology, and AI. Here we redistribute these subfields into 19 large domains according to categorization in the MAG dataset including biology, medicine, chemistry, computer science, and so on. In order to facilitate the display of the proportion, we have classified history, art, and other fields which account for a smaller proportion as “other fields”. Fig. 4 shows the distribution of various fields and the proportion of each field. Unsurprisingly, references in the fields of Biology, Medicine, and Chemistry rank in the top three, followed by computer science. Most references are related to subfields of Biology and Medicine such as drug discovery and vaccine discovery. How to combine research in Computer Science and Biomedicine has also been studied in depth [21–24].

3.3. Topic discovery and variation

In this section, we analyze the topics of all publications related to COVID-19 from various aspects, including the number of times

Table 1
Proportion of papers in different fields per week.

Time period	Business	Psychology	History	Medicine	Geology	Physics	Geography	Biology	Environmental science	Philosophy
01-07	0.00%	4.35%	8.70%	21.74%	0.00%	4.35%	13.04%	26.09%	0.00%	0.00%
01-14	0.00%	4.35%	8.70%	21.74%	0.00%	4.35%	13.04%	26.09%	0.00%	0.00%
01-21	0.00%	3.70%	7.41%	22.22%	0.00%	7.41%	11.11%	25.93%	0.00%	0.00%
01-28	0.00%	0.99%	6.93%	20.79%	3.96%	6.93%	7.92%	22.77%	0.00%	0.99%
02-04	0.00%	3.73%	6.22%	20.75%	3.32%	5.39%	6.64%	21.58%	0.00%	0.83%
02-11	0.30%	3.26%	5.34%	21.36%	2.67%	5.64%	6.53%	22.85%	0.00%	1.48%
02-18	0.45%	4.15%	6.23%	20.18%	1.93%	5.49%	7.27%	21.36%	0.00%	2.23%
02-25	0.56%	4.38%	5.90%	20.24%	1.51%	5.10%	6.85%	20.88%	0.24%	2.15%
03-03	0.84%	4.40%	5.63%	20.70%	1.38%	4.79%	6.42%	20.80%	0.30%	2.08%
03-10	1.00%	5.02%	5.36%	20.96%	1.31%	4.50%	6.16%	20.79%	0.24%	2.11%
03-17	1.20%	5.20%	4.95%	20.86%	1.47%	4.34%	5.74%	20.64%	0.22%	2.28%
03-24	1.14%	5.31%	4.68%	19.87%	1.31%	4.35%	5.70%	20.15%	0.31%	2.49%
03-31	1.36%	5.35%	4.23%	18.87%	1.37%	4.41%	5.42%	19.66%	0.35%	2.93%
04-07	1.55%	5.74%	3.76%	18.03%	1.38%	4.32%	5.10%	19.03%	0.45%	3.02%
04-14	1.61%	5.83%	3.54%	17.88%	1.34%	4.17%	4.96%	18.98%	0.46%	3.14%
04-21	1.70%	6.06%	3.28%	17.99%	1.29%	4.13%	4.80%	18.97%	0.54%	3.25%
04-28	1.70%	6.06%	3.27%	18.00%	1.29%	4.12%	4.79%	18.98%	0.55%	3.29%
05-05	1.71%	6.04%	3.27%	18.03%	1.30%	4.10%	4.79%	18.98%	0.55%	3.29%
05-12	1.71%	6.04%	3.28%	18.03%	1.30%	4.10%	4.79%	18.98%	0.55%	3.29%

Time period	Materials science	Political science	Sociology	Engineering	Economics	Mathematics	Chemistry	Computer science	Art
01-07	0.00%	8.70%	0.00%	4.35%	8.70%	0.00%	0.00%	0.00%	0.00%
01-14	0.00%	8.70%	0.00%	4.35%	8.70%	0.00%	0.00%	0.00%	0.00%
01-21	0.00%	7.41%	0.00%	3.70%	7.41%	0.00%	3.70%	0.00%	0.00%
01-28	0.00%	6.93%	0.99%	1.98%	3.96%	4.95%	7.92%	1.98%	0.00%
02-04	0.00%	6.64%	3.73%	0.83%	3.32%	3.73%	10.37%	2.90%	0.00%
02-11	0.00%	5.93%	3.56%	1.19%	2.97%	4.15%	9.79%	2.97%	0.00%
02-18	0.00%	7.12%	3.41%	1.04%	4.30%	4.45%	7.12%	2.97%	0.30%
02-25	0.16%	6.77%	3.98%	1.59%	4.06%	5.34%	6.53%	3.27%	0.48%
03-03	0.35%	7.02%	3.80%	1.58%	4.35%	4.89%	6.37%	3.66%	0.64%
03-10	0.42%	7.09%	3.74%	1.59%	4.32%	4.74%	6.43%	3.63%	0.59%
03-17	0.42%	7.18%	3.58%	1.59%	4.58%	4.49%	6.72%	3.92%	0.64%
03-24	0.37%	7.34%	3.93%	1.66%	5.11%	4.89%	6.32%	4.41%	0.66%
03-31	0.41%	7.20%	3.84%	1.86%	5.42%	5.42%	6.20%	5.03%	0.65%
04-07	0.41%	7.28%	3.98%	2.18%	5.88%	5.85%	5.90%	5.58%	0.57%
04-14	0.46%	7.28%	3.95%	2.22%	5.98%	5.84%	5.97%	5.76%	0.61%
04-21	0.46%	7.20%	3.86%	2.48%	6.10%	5.72%	5.87%	5.65%	0.63%
04-28	0.45%	7.20%	3.85%	2.47%	6.11%	5.71%	5.87%	5.65%	0.63%
05-05	0.46%	7.21%	3.81%	2.47%	6.09%	5.70%	5.91%	5.65%	0.64%
05-12	0.46%	7.22%	3.81%	2.46%	6.10%	5.69%	5.91%	5.65%	0.64%

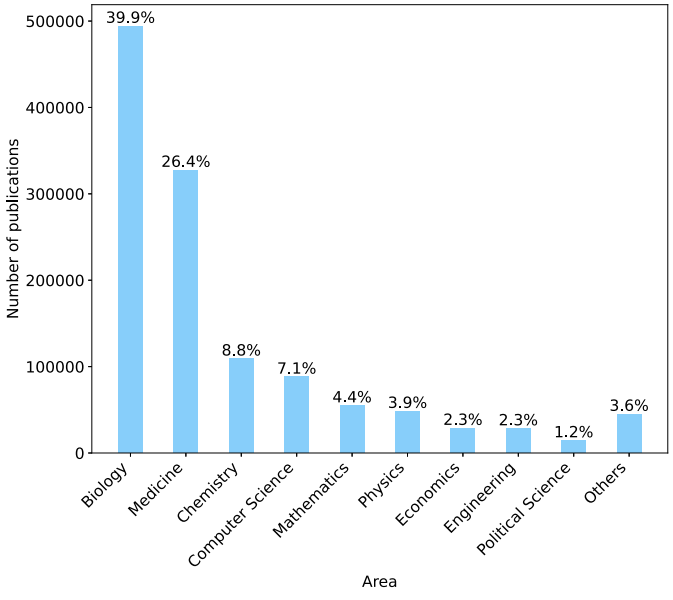


Fig. 4. Area distribution of references in COVID-19 related research.

the topics appears, the variation of topics over time, and the number of topic co-occurrences.

3.3.1. Topic discovery

After performing the LDA model on abstracts of all papers, we get two posterior distributions: the posterior topic distribution θ_m of each paper d_m and the posterior word distribution φ_k of each topic z_k. θ_{mk} can be deemed as the proportion of topic z_k in document d_m. We represent φ_k of each topic z_k as the word-cloud in Fig. 5 and Fig. 6.

Due to space limit, we only show the first 20 words with the highest posterior probability in each topic. In the word-cloud of each topic, the size of each word is directly proportional to its posterior probability. The topics are inferred from the distribution of words in the topic and related knowledge. The results can be used to analyze the topics of COVID-19 related papers because they correspond well with the research fields. For example, Topic 0: “probability, pregnant, non-pregnant, women, participants, UK, outbreaks, HCW, US, adults, ...” is mainly related to the impact of COVID-19 on pregnancy. Topic 1: “travel, travelers, aerosols, imported, liver, ARBs, united, states, locations, countermeasures, ...” mainly focuses on the impact of travel. In addition, there are some general topics related to the academic database. For example, Topic 27: “review, published, 25-th, PubMed, systematic, literature,

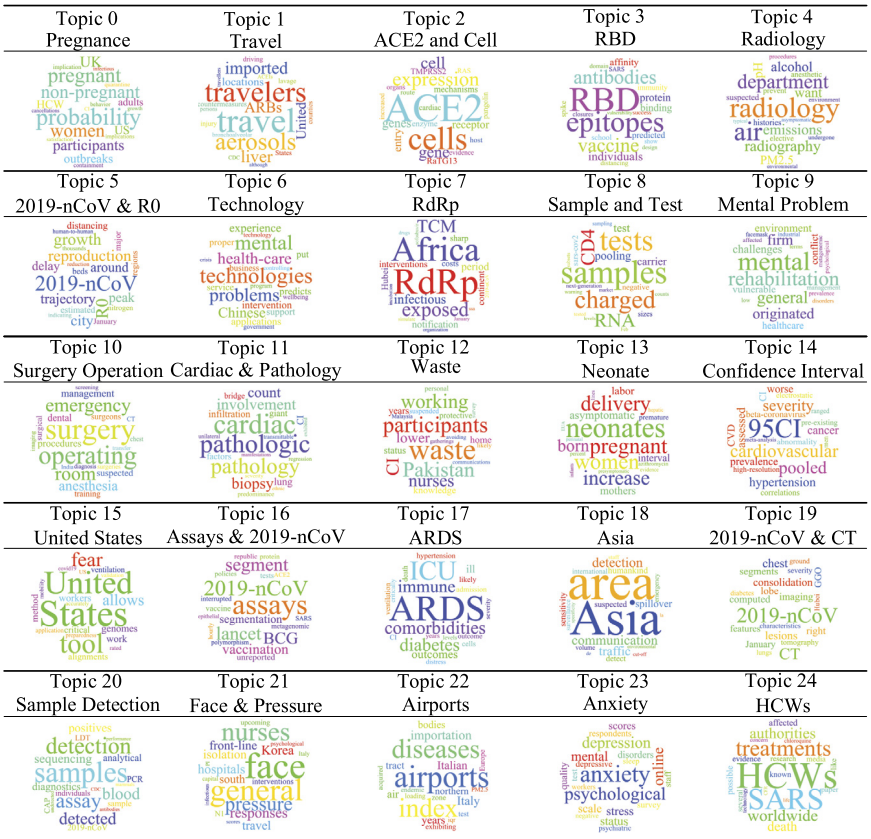


Fig. 5. Word-cloud of Topic 0 – Topic 24.

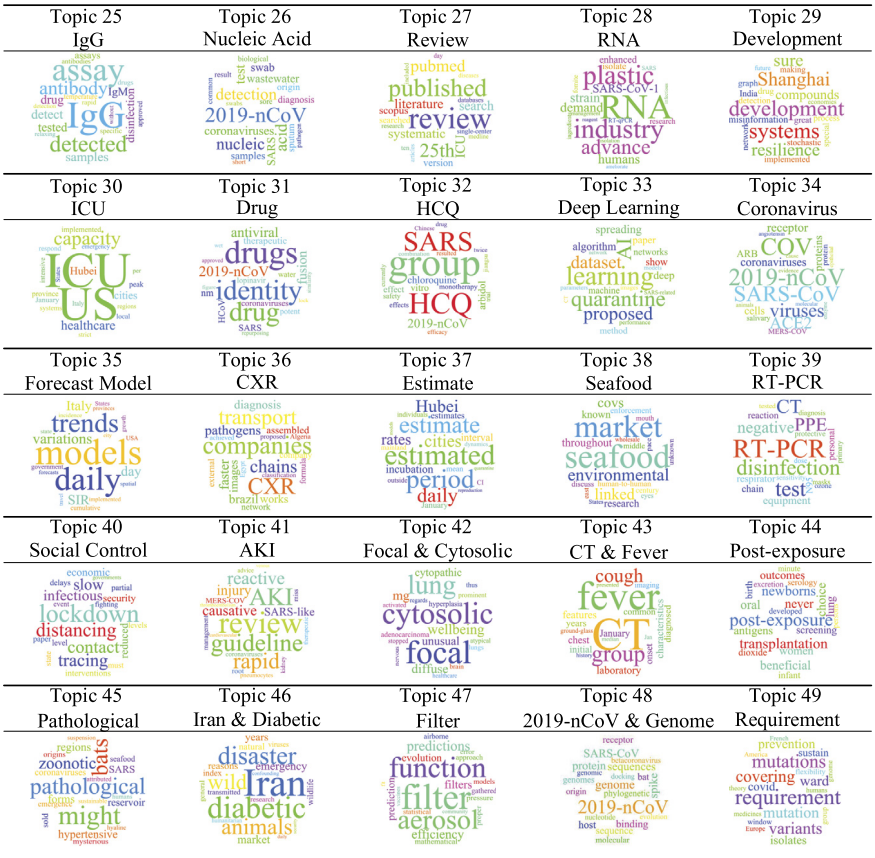


Fig. 6. Word-cloud of Topic 25 – Topic 49.

Table 2
r_k and c_k for each topic.

Topic	r_k	$r_k^{(Feb)}$	$r_k^{(Mar)}$	$r_k^{(Apr)}$	$r_k^{(May)}$	c_k	Topic	r_k	$r_k^{(Feb)}$	$r_k^{(Mar)}$	$r_k^{(Apr)}$	$r_k^{(May)}$	c_k
14	38.35	33.06	3.84	1.07	0.37	55.64	42	14.89	7.71	6.04	0.92	0.22	38.63
46	35.82	30.06	4.65	0.95	0.16	65.97	21	14.73	9.99	3.7	0.91	0.14	100.17
27	33.01	26.88	4.58	1.30	0.25	94.56	3	14.06	9.42	3.02	1.44	0.18	96.33
40	26.25	19.10	5.14	1.80	0.21	146.86	32	13.54	8.25	4.15	0.88	0.27	112.73
44	25.37	18.74	5.35	1.01	0.27	73.20	6	12.92	4.9	6.93	0.91	0.18	73.37
47	22.05	14.95	5.45	1.44	0.21	103.72	20	12.76	7.63	3.38	1.52	0.22	83.74
12	20.87	14.48	4.83	1.28	0.28	125.38	36	12.55	5.83	5.07	1.4	0.26	60.09
25	19.92	12.72	5.56	1.40	0.24	165.33	2	12.5	7.68	3.36	1.28	0.18	159.53
13	19.14	13.31	4.34	1.27	0.23	67.78	28	12.5	5.6	5.4	1.28	0.21	76.26
39	18.32	12.09	4.47	1.47	0.29	116.82	9	12.42	5.95	5.27	1.0	0.2	114.45
7	18.01	13.96	2.98	0.86	0.21	68.59	45	12.15	8.98	1.84	1.14	0.19	49.73
4	17.46	7.01	9.29	0.95	0.21	61.34	1	12.08	7.93	3.16	0.85	0.14	84.64
0	17.35	10.94	4.82	1.36	0.23	79.16	30	12.03	6.56	3.91	1.36	0.21	247.30
22	16.91	12.36	3.21	1.02	0.32	67.74	37	11.82	8.04	2.07	1.51	0.2	535.14
23	16.71	11.55	3.63	1.15	0.38	86.69	24	11.7	5.68	4.82	1.04	0.17	225.69
35	16.69	10.79	4.06	1.66	0.18	234.07	29	11.47	5.54	4.3	1.43	0.2	148.58
8	16.28	11.06	3.67	1.25	0.3	63.72	38	11.44	5.93	4.4	0.98	0.13	85.61
33	16.01	10.42	3.68	1.74	0.17	247.64	26	9.39	6.01	2.15	1.04	0.19	128.84
11	15.94	11.38	3.07	1.29	0.21	79.58	18	9.35	5.12	2.87	1.19	0.17	52.93
49	15.94	7.87	6.63	1.22	0.22	72.31	31	9.34	4.95	3.22	1.03	0.14	24.26
15	15.72	6.85	7.06	1.63	0.18	114.28	34	9.31	4.66	3.54	0.9	0.21	111.09
10	15.56	6.44	8.17	0.81	0.13	102.86	16	9.06	4.32	3.13	1.41	0.2	81.33
17	15.53	9.18	5.08	0.99	0.27	224.88	41	8.79	3.21	4.48	0.89	0.22	109.96
43	15.38	11.65	2.81	0.64	0.29	315.02	48	7.97	4.75	2.06	1.0	0.15	166.28
19	15.03	11.73	2.05	1.0	0.24	104.01	5	7.56	3.65	3.12	0.67	0.12	101.02

search, ICU, scopus, version, ..." involves many words in academic papers.

3.3.2. Topic proportion and topic distribution over time

Table 2 lists r_k of all topics in a decreasing order. $r_k^{m-1} > 1$ indicates that the proportion of topic z_k has increased from the month m to the month $m-1$. These topics can be regarded as hot topics. On the contrary, $r_k^{m-1} < 1$ indicates that the proportion of topic z_k has decreased. It can be seen from the results that the trend of each topic is different every month. Concerning the value of r_k, the top 5 hottest topics are: Topic 14: "Confidence Interval: 95CI, cardiovascular, pooled, severity, ...", Topic 46: "Iran & Diabetic: Iran, diabetic, disaster, animals, ...", Topic 27: Review: "review, published, pubmed, systematic, ...", Topic 40: "Social Control: lockdown, distancing, contact, tracing, ...", Topic 44: "Post-exposure: post-exposure, transplantation, newborns, choice, ...". From the evolution of these topics over time (Fig. 8(a)), we can easily find that the percentage of papers containing these topics has been increasing over time. These hot topics are mainly related to the treatment of comorbid symptoms of COVID-19 and other diseases, as well as the social control of COVID-19. This shows that among all the hot topics of COVID-19, the treatment and control of disease are the most widely concerned. Scholars pay attention to Biology and Medicine related research including epidemic control and its related symptoms as well as its treatment options.

The cumulative proportion c_k of topics is also presented in Table 2. In addition, we also count the number of times each topic appears in the paper. We can regard the topic with a larger cumulative proportion as a more popular topic. The five most popular topics are: Topic 37, Topic 43, Topic 33, Topic 30, and Topic 35. These topics are mainly about the estimation and prediction of COVID-19 transmission (Topic 37), diagnosis/confirmation (Topic 43 and Topic 30), and related applications of deep learning technology (Topic 33 and Topic 35). The results not only reveal the focus that scholars pay attention to in the COVID-19 research, but also highlight the need for AI in future pandemics.

The distribution of all topics is presented in Fig. 7. Topics (from 0 to 49) are arranged from bottom to top. The horizontal axis is a time period divided every 7 days. The length of different colors on the vertical axis indicates the proportion of different topics. Fig. 7 illustrates the trend of different topics over time.

Table 3
Top 20 topic pairs with the largest co-occurrence times.

Rank	Co-occurrence times	Topic pairs	Rank	Co-occurrence times	Topic pairs
1	759	(35, 37)	11	433	(35, 40)
2	700	(30, 37)	12	427	(29, 33)
3	579	(37, 40)	13	417	(24, 37)
4	575	(33, 37)	14	414	(24, 30)
5	497	(30, 35)	15	412	(29, 35)
6	484	(33, 35)	16	409	(37, 47)
7	461	(29, 37)	17	403	(12, 37)
8	459	(30, 40)	18	392	(30, 33)
9	443	(5, 37)	19	379	(15, 37)
10	434	(17, 43)	20	374	(0, 37) (33, 40)

We also analyze the trends of the above five most popular topics (Fig. 8(b)). We find that Topic 37 accounts for a large proportion from mid-January to mid-February, and Topic 43 accounts for a large proportion from the late February to mid-March. The proportion of Topic 33, 30, and 35 is relatively stable throughout the whole period. In the early stage of the epidemic, scholars pay attention to the estimation of the epidemic and the diagnosis of COVID-19. Predictions of epidemic outbreaks can help decision makers formulate future measures to prevent epidemics. With the development of computer science technology, especially deep learning technology, many aspects of the COVID-19 crisis at different scales including molecular, drug discovery, and societal applications are benefited [25].

3.3.3. Co-occurrence analysis

A paper may contain multiple topics. There may be strong connections between different topics that frequently appear together in the same paper. To explore the interaction among these topics, we calculate and analyze the co-occurrence between topics in all papers. Specially, we calculate the probability of two topics appearing in the same paper and visualize results using heat map (see Fig. 9).

We establish a co-occurrence matrix $C \in \mathbb{R}^{50\times 50}$, where C_{ij} is the co-occurrence times between Topic z_i and Topic z_j in all papers, $C_{ij} = C_{ji}$, and $C_{ij} = 0$ if $i = j$. In Fig. 9, the darker the color is, the smaller the co-occurrence times of two topics are. We find

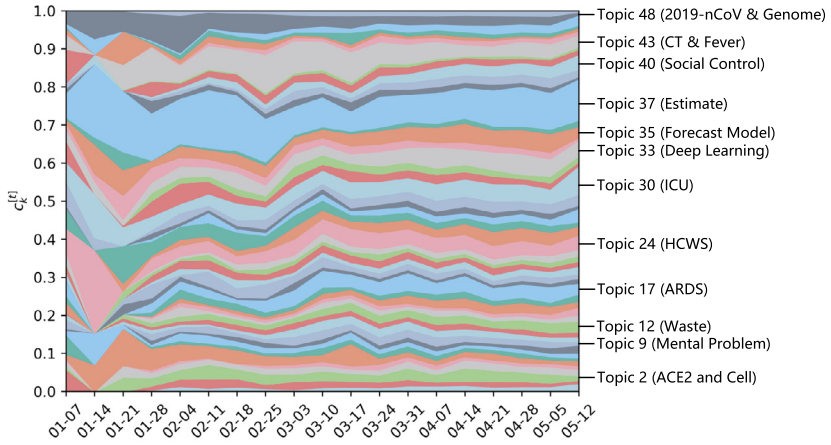
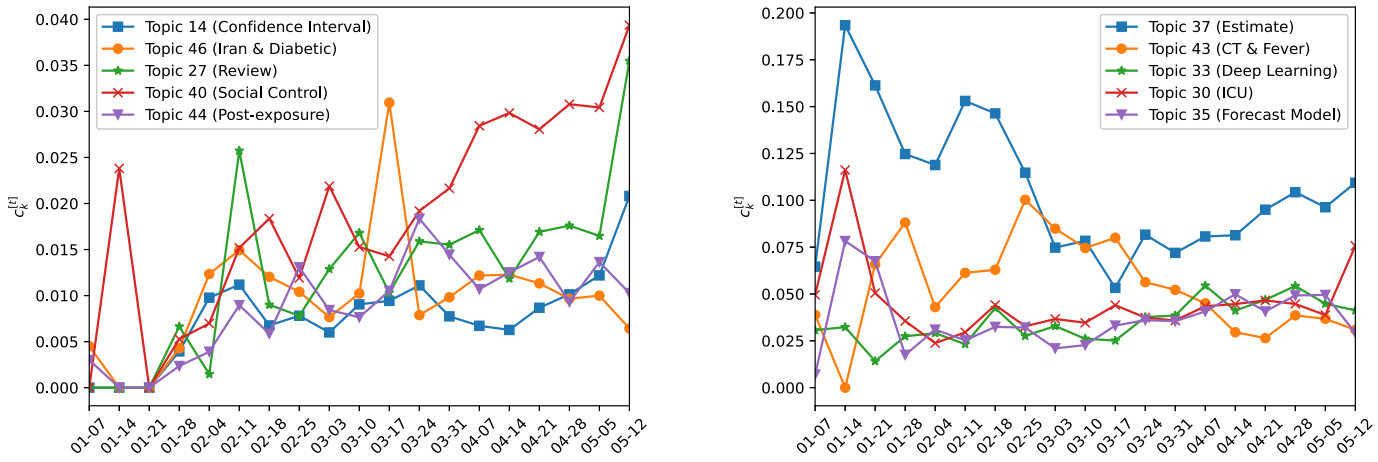


Fig. 7. Topic variation over time.



(a) Top five hottest topics (b) Top five most popular topics

Fig. 8. Topic variation trends over time of top five hottest topics and top five most popular topics.

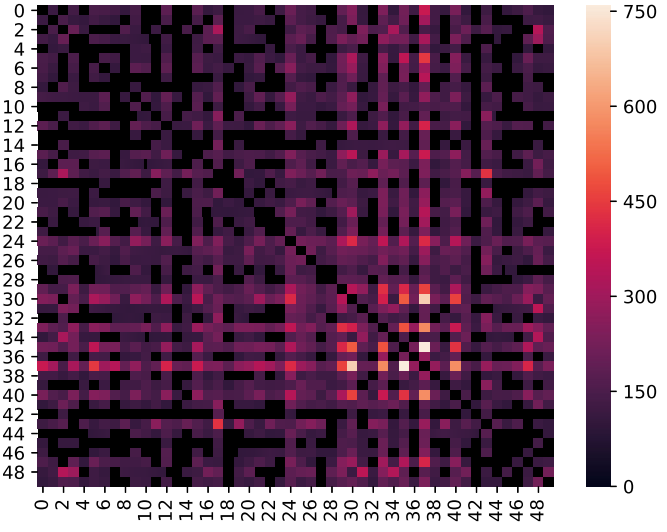


Fig. 9. Co-occurrence probability for topics in the study period. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

that the elements at other positions are not zero except for the diagonal elements of matrix C. This means that any two topics have appeared in the same paper. We can see that the co-occurrence times of most topics are below 300.

In order to better find topics with strong correlation, we list the top 20 topic pairs with the largest co-occurrence times in Table 3. The top five topics for the number of co-occurrences are Topic 35: "models, daily, trends, variations, Italy, SIR, day, government, cumulative, implemented, ..." and Topic 37: "estimated, period, estimate, daily, Hubei, cities, rates, incubation, interval, estimates, ...", Topic 30: "ICU, US, capacity, healthcare, cities, Hubei, implemented, province, January, peak, ..." and Topic 37, Topic 37 and Topic 40: "lockdown, distancing, contact, tracing, slow, infectious, economic, reduced, security, interventions, ...", Topic 33: "learning, quarantine, AI, proposed, dataset, spreading, algorithm, deep, show, paper, ..." and Topic 37, Topic 30 and Topic 35. Strong correlations between these topics are observed.

3.4. Innovative topics and papers discovery

To identify COVID-19 related research breakthroughs, we predict the innovative papers based on the approach mentioned in Section 2.2. To obtain deep insights into innovative papers, we select ten papers to analyze the field, topics, and keywords in the papers (as shown in Table 4). Note that all of these ten papers are published during January and March in 2020 with the development of the pandemic.

It is observed that the majority of COVID-19 papers belong to Medicine. Others fall into subfields such as Internal Medicine, Biochemistry, and Psychiatry. Three of them are about the clinical

Table 4
The top ten papers with the most innovation according to innovative paper identification algorithm.

Title	Published time
COVID-19 and the anti-lessons of history	2020-03-02
The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned?	2020-02-22
A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)	2020-02-06
Communicating the risk of death from novel coronavirus disease (COVID-19)	2020-02-21
Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020	2020-02-14
COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures	2020-02-28
Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study	2020-02-20
Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak	2020-02-07
Eco-virological preliminary study of potentially emerging pathogens in Hedgehogs (Erinaceus europaeus) recovered at a wildlife treatment and rehabilitation center in northern Italy	2020-03-01
Clinical characteristics of 2019 novel coronavirus infection in China	2020-02-09

Table 5
Number of publications based on country analysis.

Country	The number of publications before March 11	The number of publications after March 11
China	521	640
USA	19	77
Italy	17	174
Brazil	6	21
Spain	2	59
Germany	10	46
Korea	39	86
France	10	42
Iran	30	47
Japan	10	53

Table 6
Keywords distribution in innovative topics.

Word	Number	Word	Number	Word	Number
January	8	Pregnant	1	Technologies	1
Travel	6	Post-Exposure	1	States	1
2019-nCoV	6	Outbreaks	1	Spain	1
Estimate	5	Nurses	1	Would	1
Growth	3	Non-Pregnant	1	Samples	1
Hubei	3	Months	1	Regions	1
SARS	2	Mental	1	Reached	1
Research	2	Mainland	1	Quarantine	1
Reproduction	2	Italy	1	Psychological	1
Province	2	Isolation	1	Protein	1
Proper	2	Interventions	1	Communication	1
Iran	2	Importation	1	Cities	1
Interval	2	Healthcare	1	Probability	1
International	2	Hangzhou	1	Chest	1
ICU	2	Factors	1	Cancer	1
Fever	2	Experience	1	Beijing	1
CT	2	Emergency	1	Basic	1
ARDS	2	Drugs	1	Asymptomatic	1
Zhejiang	1	Distancing	1	Anxiety	1
Years	1	Disaster	1	Ace2	1
Women	1	Death	1	Trends	1
Daily	1				

characteristics and clinical guidance. Five papers focus on the epidemic transmission and outbreak estimation. Two papers compare COVID-19 with SARS-CoV to gain experience from SARS-CoV research. Nearly 61% of papers are based on China's data for analysis before March 11. The pandemic situation in other countries such as the United States and Italy are also mentioned in papers published later [26–28] (see Table 5).

Up to May 2020, apart from social distancing there are neither other effective measures to prevent infection of the pandemic, nor effective drugs or vaccines to treat the virus. Therefore, before the pandemic is completely controlled, the treatment and prevention of the disease are expected to remain active research topics of COVID-19. However, with the extension of the pandemic duration, emerging topics such as the impact to psychology and sociology are expected to arise.

Table 6 lists the distribution of the keywords in innovative topics. In these papers, "January" is the most frequent word, followed by "Travel". The topics with a lower frequency are "2019-nCoV, Hubei, Growth, and Estimated". If "January" and "Travel" are words that every paper will mention, these four words are more like the target of each paper, as five papers focus on the epidemic transmission and outbreak estimation. The words that often appear are "SARS-CoV, ARDS, and Fever". The comparison between COVID-19 and SARS-CoV is also the focus of researchers. They hope to gain experience from SARS-CoV research to help with the cure and detection of COVID-19. Other emerging words like "CT, psychology" appear once or twice. Some papers may focus on the effect of CT or the mental impact of the epidemic. In papers about clinical characteristics and virus detection, low-frequency words such as "ICU" will appear. In papers about candidates, words like drug and therapeutic will appear. In papers about prevention and infection, words like mobility and chest will appear. Therefore, we can draw a conclusion from these words in innovative topics. In the fields of Medicine and Chemistry, words about cure and infection will appear. Apart from these two fields, researchers from other fields have also contributed to COVID-19 research in their own way.

4. Conclusion

The purpose of this paper is to explore the anatomy of early COVID-19 related research, which can provide the society with a clearer understanding of the epidemic and help combat the COVID-19 pandemic at the early stage. By leveraging the methodology of AI and big data, this paper shows the scientific contribution against COVID-19 in different domains. Specifically, we have highlighted evolutionary patterns along the line of focus topics based on scientometric analysis. Our results provide important implications for understanding how research communities contribute to the current situation. The major findings include:

- In the early days of the epidemic, the majority of COVID-19 research focused on Medicine and Biology, such as coronavirus mechanisms and drug discovery. With the development of the epidemic, the research focus of COVID-19 has expanded to other research hotspots, such as analyzing the impact of coronavirus from social, economic, and psychological perspectives. In addition, AI is always a hot topic, which is also regarded as a common and effective method to be conducted in various research of different fields. In the studying process, we observe

that topics are not developing independently but interrelating with each other.

- In terms of citation behavior, results show that the COVID-19 cases are highly correlated with the SARS pandemic. Scholars learn from relevant research on SARS to find strategies for infection prevention or patient treatment of COVID-19. Similarly, most references are published in 2010-2019, which can be explained by the outbreak of the Middle East Respiratory Syndrome (MERS) in 2012. Besides, the proportion of references from computer science ranks forth, following Biology, Medicine, and Chemistry, which indicates the importance of computer science in academic research.
- Through distinguishing innovative papers and keywords, we find that signposted research of COVID-19 related research mainly focus on clinical characteristics and virus detection. The treatment and prevention of the virus will still be hot issues of COVID-19 before the pandemic is completely under control. In the analysis of 50 key topics, experimental results also indicate that the need for AI is highlighted in the future academic research of pandemics.

Overall, this study provides readers with deeper insights into topic development and the evolution of COVID-19 research at the early stage. It helps researchers understand how academic communities deal with the on-going pandemic, how they respond to the evolution of the current pandemic, and what they do to combat future pandemics. Furthermore, the temporal variation of topics could help researchers understand the research trends to guide subsequent research directions. We acknowledge the difficulty of contributing through academic research in the current situation. Nonetheless, we hope this paper could help research communities understand the value of academic research and the promising domains for collaboration, with the ultimate goal to direct research for actions to prevent and to treat pandemic.

There are also some minor limitations in this research. First of all, the literature data used in this paper is derived from CORD-19, without considering other databases. In the future, we may focus on more complete datasets for more comprehensive bibliometric results. Apart from this, it is also interesting to analyze COVID-19 related research from other aspects such as co-authorship network analysis to gain further insights.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. West, S. Michie, G.J. Rubin, R. Amlôt, Applying principles of behaviour change to reduce SARS-CoV-2 transmission, Nat. Hum. Behav. 4 (2020) 451–459.

[2] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of COVID-19, JAMA 323 (14) (2020) 1406–1407.

[3] J.T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P.M. de Salazar, B.J. Cowling, M. Lipsitch, G.M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China, Nat. Med. 26 (2020) 506–510.

[4] J. Zhang, M. Litvinova, Y. Liang, Y. Wang, W. Wang, S. Zhao, Q. Wu, S. Merler, C. Viboud, A. Vespignani, et al., Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China, Science 368 (6498) (2020) 1481–1486.

[5] H. Rossman, A. Keshet, S. Shilo, A. Gavrieli, T. Bauman, O. Cohen, E. Shelly, R. Balicer, B. Geiger, Y. Dor, et al., A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys, Nat. Med. 26 (5) (2020) 634–638.

[6] L. Dong, S. Hu, J. Gao, Discovering drugs to treat coronavirus disease 2019 (COVID-19), Drug. Discov. Ther. 14 (1) (2020) 58–60.

[7] T. Chookajorn, Evolving COVID-19 conundrum and its impact, Proc. Natl. Acad. Sci. 117 (23) (2020) 12520–12521.

[8] L. Sun, Y. Yin, Discovering themes and trends in transportation research using topic modeling, Transp. Res., Part C, Emerg. Technol. 77 (2017) 49–66.

[9] J. Liu, J. Tian, X. Kong, I. Lee, F. Xia, Two decades of information systems: a bibliometric review, Scientometrics 118 (2) (2019) 617–643.

[10] F. Xia, W. Wang, T.M. Bekele, H. Liu, Big scholarly data: a survey, IEEE Trans. Big Data 3 (1) (2017) 18–35.

[11] H. Jang, E. Rempel, D. Roth, G. Carenini, N.Z. Janjua, Tracking Covid-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis, J. Med. Internet Res. 23 (2) (2021) e25431.

[12] C. Ordun, S. Purushotham, E. Raff, Exploratory analysis of COVID-19 tweets using topic modeling, umap, and digraphs, preprint, arXiv:2005.03082, 2020.

[13] S.K. Sonbhadra, S. Agarwal, P. Nagabhushan, Target specific mining of COVID-19 scholarly articles using one-class approach, Chaos Solitons Fractals 140 (2020) 110155.

[14] M. Dong, X. Cao, M. Liang, L. Li, H. Liang, G. Liu, Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling, medRxiv, 2020.

[15] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al., Cord-19: the Covid-19 open research dataset, preprint, arXiv:2004.10706, 2020.

[16] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, K. Wang, An overview of Microsoft Academic Service (MAS) and applications, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 243–246.

[17] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[18] P. Savov, A. Jatowt, R. Nielek, Towards understanding the evolution of the WWW conference, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 835–836.

[19] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, R. Agha, The socio-economic implications of the coronavirus pandemic (COVID-19): a review, Int. J. Surg. 78 (2020) 185–193.

[20] B. Gavin, J. Lyne, F. McNicholas, Mental health and the COVID-19 pandemic, Ir. J. Psychol. Med. (2020) 1–7.

[21] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710.

[22] L. Heo, M. Feig, Modeling of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins by machine learning and physics-based refinement, bioRxiv, 2020.

[23] J. Fauqueur, A. Thillaisundara, T. Togia, Constructing large scale biomedical knowledge bases from scratch with rapid annotation of interpretable patterns, preprint, arXiv:1907.01417, 2019.

[24] B.R. Beck, B. Shin, Y. Choi, S. Park, K. Kang, Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model, Comput. Struct. Biotechnol. J. 18 (2020) 784–790.

[25] J. Bullock, K.H. Pham, C.S.N. Lam, M. Luengo-Oroz, et al., Mapping the landscape of artificial intelligence applications against COVID-19, preprint, arXiv: 2003.11336, 2020.

[26] A. Remuzzi, G. Remuzzi, Covid-19 and Italy: what next?, Lancet 395 (10231) (2020) 1225–1228.

[27] G. Onder, G. Rezza, S. Brusaferro, Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy, JAMA 323 (18) (2020) 1775–1776.

[28] S.M. Moghadas, A. Shoukat, M.C. Fitzpatrick, C.R. Wells, P. Sah, A. Pandey, J.D. Sachs, Z. Wang, L.A. Meyers, B.H. Singer, et al., Projecting hospital utilization during the COVID-19 outbreaks in the United States, Proc. Natl. Acad. Sci. 117 (16) (2020) 9122–9126.