


# Finding answers to COVID-19-specific questions: An information retrieval system based on latent keywords and adapted TF-IDF

Journal of Information Science  
1–17  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/01655515221110995  
journals.sagepub.com/home/jis  


**Jorge Chamorro-Padial**   
CITIC-UGR, Universidad de Granada, Spain

**Francisco-Javier Rodrigo-Ginés**   
NLP & IR Group, UNED, Spain

**Rosa Rodríguez-Sánchez**  
Departamento de Ciencias de la Computación e Inteligencia Artificial, CITIC-UGR, Universidad de Granada, Spain

## Abstract

The scientific community has reacted to the COVID-19 outbreak by producing a high number of literary works that are helping us to understand a variety of topics related to the pandemic from different perspectives. Dealing with this large amount of information can be challenging, especially when researchers need to find answers to complex questions about specific topics. We present an Information Retrieval System that uses latent information to select relevant works related to specific concepts. By applying Latent Dirichlet Allocation (LDA) models to documents, we can identify key concepts related to a specific query and a corpus. Our method is iterative in that, from an initial input query defined by the user, the original query is expanded for each subsequent iteration. In addition, our method is able to work with a limited amount of information per article. We have tested the performance of our proposal using human validation and two evaluation strategies, achieving good results in both of them. Concerning the first strategy, we performed two surveys to determine the performance of our model. For all the categories that were studied, precision was always greater than 0.6, while accuracy was always greater than 0.8. The second strategy also showed good results, achieving a precision of 1.0 for one category and scoring over 0.7 points overall.

## Keywords

ATFIDF; COVID-19; document filtering; information retrieval; keywords generation; latent Dirichlet allocation; TF-IDF

## 1. Introduction

Coronavirus disease (also called COVID-19) was first detected in Wuhan, in the Hubei province of China and was reported to the World Health Organization (WHO) by the Chinese government on 31 December 2019 [1]. The illness has been spreading since the beginning of 2020 and officially became a pandemic in March when the virus had already infected more than 150,000 people worldwide [2,3].

Since the beginning of 2020 and due to the social, economic and political repercussions generated by the evolution of the virus [4,5], the media and the scientific community have published an incredible amount of information on the aforementioned disease [6]. In just 3 months from the initial notification of the disease by the Chinese authorities, a total of 1596 publications had been generated about COVID-19, 66% of them from China, while in April, the number of publications rose to more than 6500 [7,8]. At a social level, the pandemic has also attracted the attention of social network

---

### Corresponding author:

Francisco-Javier Rodrigo-Ginés, NLP & IR Group, UNED, 28040 Madrid, Spain.  
Email: frodrigo@invi.uned.es

users [5]. In these circumstances, the excess of information and the lack of knowledge about the pandemic have led to the spread of inaccurate or false information (also known as *fake news*).

In the current environment, it is necessary and relevant to establish strategies to analyse the large amount of information that is continuously being generated in the scientific world in order to make the task of responding to the different needs that arise from different fields easier. Such fields include health, social, economics, ethics, educational and political, to name a few.

In this article, we propose a method to extract scientific literature based on a variety of topics. Our method facilitates the work of identifying which works written about COVID-19 could help respond to certain questions. Furthermore, our method also makes it possible to differentiate between articles that do and do not address purely health-related issues. This article aims to make it easier for researchers to find and retrieve scientific literature related to complex or abstract topics, thus making it easier to find answers to complicated questions and to provide them with a complementary method for using information retrieval systems based on standalone keywords.

To do this, we worked with a dataset of articles on COVID-19 and other related areas. We extracted the different topics for each article, analysing its title and abstract. Then, we filtered the documents by performing a coincidence analysis on the terms of the topics with the terms of the query.

We propose a method to find answers to complex and abstract questions by exploring latent concepts hidden in the titles and abstracts of scientific works. Our main goal was to respond to the call to action from the White House and other various research groups. They had prepared a dataset of COVID-19-related articles so as to solve urgent and relevant problems related to the pandemic [9]. While plenty of proposals have been offered by different authors (some of them are mentioned in the following section), we wanted to design a simpler model that would be able to find answers on COVID-19 adequately. More importantly, humans have validated our method, while most authors use automatic systems to measure or benchmark their proposals.

While our method has been tested specifically with COVID-19 data, it can be generalised and used in different areas. Our work is structured as follows:

- The ‘State of the art’ section reviews the latest published studies, mainly in the field of bibliometrics and natural language processing (NLP), as related to the COVID-19 disease.
- The ‘Methodology’ section describes the dataset used and the proposed method.
- The ‘Results’ section shows the results obtained by our method.
- Finally, the ‘Conclusion’ section states the main findings of our work.

## 2. State of the art

Since the outbreak of the pandemic, there have been many efforts to analyse the behaviour of the scientific community through Bibliometry. Lou et al. [10] and Nasab and Rahim [11] analyse the relationships between authors and highly cited articles about both the COVID-19 disease and the SARS-CoV-2 virus, as well as the number of papers produced by country, providing empirical data that proves the growing interest that the pandemic has had among not only authors, but also journals of very high impact. In both papers, scientific publications are also broken down by different subject areas, with epidemiology and virology being the fields that have received the greatest number of publications. Although [3] also performs a bibliometric analysis where they reached similar conclusions as [10,11], the authors also provided an interpretation for the significant differences in publications observed by country, based solely on the gross domestic product (GDP) and the number of inhabitants. A high level of saturation in the healthcare system can have a negative impact on the number of publications in a country. This is the case, for example, in Italy.

This massive amount of information about the pandemic requires useful strategies that can help facilitate the scientific community in finding the desired information, while weeding out those works that may not be relevant for them. The large number of pre-existing works in the highly specific fields in the world of Medicine [3,10,11] also makes it more difficult to find information on certain less well-studied areas, such as education or ethics.

In addition, the evolution of the pandemic has unleashed plenty of social repercussions that should be studied and taken into consideration, for example, the vaccine opposition and COVID-19 denial movements which have had a considerable impact on social networks [12,13]. Work-from-home has also become a trend during the outbreak and is paving the way for an important transformation in terms of labour relations [14]. The pandemic has also had repercussions on cities, where marginalised populations are receiving a disproportional impact on their health and well-being. Urban planning is also learning and growing from this situation as well [15].

Lou et al. [10] highlight the limitations of their bibliometric analysis under the current conditions, where a large number of articles on the subject of COVID-19 are being published, in different languages. Being such a current event, it is impossible to draw firm conclusions in a field so dependent on current events.

In order to facilitate the task of obtaining relevant information for researchers, it is possible to apply techniques from the field of NLP. IBM, for example, offers a service to extract relevant content from a corpus of articles about COVID-19, allowing researchers to ask specific questions about COVID-19 and analyse the related existing information.<sup>1</sup>

NLP is also being used to analyse the social interactions caused by the pandemic. Cinelli et al. [16] extract topics from a corpus of text from different social networks and perform a discourse analysis to infer key concepts that have been used by the users of social media and the patterns of information dissemination. Lopez et al. [17] analyze Twitter with text mining techniques in order to conduct a multi-language analysis of the speech. In Singh et al. [18] an analysis of comments on Twitter about the pandemic was carried out but, this time, by analysing the dissemination of truthful information and misinformation through the social network. In Schild et al. [19] the authors analysed the phenomenon of sinophobia on 4chan and Twitter in relation to the pandemic through word embeddings, linking different news about Donald Trump, the WHO and the Chinese Government.

With the aim of helping researchers apply NLP in the search for information concerning the pandemic, Riloff et al. [20] have designed a toolbox that includes a set of English dictionaries with relevant concepts related to COVID-19. Latif et al. [21] list the different areas of study where the use of artificial intelligence and machine learning could be relevant, as well as a compilation of datasets, resources and initiatives carried out to improve the current knowledge about the disease.

Keyword extraction models have been widely used to classify different domains of knowledge in scientific articles [22]. We can define keywords from two different perspectives: sociocultural and statistical. Traditionally, any word that includes culturally and socially relevant concepts has been intuitively considered as a keyword [23]. From the point of view of corpus linguistics, keywords are extracted by using statistical processes, commonly comparing their frequency in the text to be analysed with their frequency in a reference corpus. Using this type of technique, three types of keywords are usually obtained: proper names, concepts that explain the content of the text, and frequent words such as pronouns and prepositions that can be used as style and not content indicators [24].

Some information retrieval systems have been created to help researchers find relevant information. Named entity recognition (NER) can be useful in collecting COVID-19 information from statements, which can be considered an alternative to document retrieval systems. Wang et al. [25] developed a web-based system to find textual evidence from COVID-19 document corpus.

CO-Search is an information retrieval system that is able to extract search queries from natural language questions and to retrieve scientific literature about COVID-19 [26]. CO-Search uses a SBERT model to create a latent space with queries and documents. CO-Search results are evaluated using TREC-COVID, an evaluation system that helps researchers find searching algorithms and information discovering methods to manage the existing literature around COVID-19 [27].

For researchers, collecting scientific literature is crucial in order to be up to date with the newest and most relevant knowledge for their research areas and to provide solid background knowledge that will allow them to effectively contribute to their field. The explosive growth of new scientific literature makes it difficult to identify suitable papers and is becoming an increasingly complex task [28].

When doing a literature review, researchers need to maximise the ‘relevance’ of the collected literature, but ‘relevance’ is not directly measurable, and some level of uncertainty is inevitable [29]. While information retrieval systems are nowadays an cornerstone for researchers, question answering systems are becoming a powerful tool that may help to find more relevant knowledge [26,30].

### 3. Methodology

#### 3.1. Dataset description

In March 2020, due to the COVID-19 global pandemic, the *Allen Institute for AI* coordinated by the *White House Office of Science and Technology Policy* and in association with several initiatives published CORD-19,<sup>2</sup> an open dataset of over 50,000 papers on COVID-19, SARS-CoV-2, coronavirus and other related study areas.

The idea behind the publication of this dataset was that after the increase in the academic literature on COVID-19 [8], the computer science community could apply text mining and natural language processing methods in order to digest and retrieve significant information and provide it to the medical research community.

The CORD-19 dataset contains multilingual information, but we have only dealt with English papers in this study. Thus, a new stage removing non-English information was applied.

After deleting non-English or duplicated papers and discarding papers without abstracts, titles, or references, we got a reduced dataset of 25,004 instances. Each instance of the dataset has the following information: title, authors, abstract, body text, references and publication date.

### 3.2. Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model used to extract the latent topic structure of text documents. LDA is a machine-learning technique used in different areas such as retrieval field, document classification and topic modelling [31].

In this article, we have implemented the LDA provided by the Scikit-Learn project: a machine-learning library for the Python Programming Language.<sup>3</sup>

### 3.3. Method

In this section, we will present the main contribution of this study: an information retrieval system that allows users to extract relevant papers when given certain search terms. Our approach is based on keyword extraction using the LDA topic modelling technique on pseudo-documents generated from the papers in the dataset.

The idea of this work is to relate both concepts of keywords. By adding latent keywords in the dataset papers, we can broaden the search spectrum and obtain more relevant results. As in other Information Retrieval Systems, the user introduces input information to perform a query and obtain an output. The input information is a set of keywords containing the most important ideas that the user wishes to examine. In our method and thanks to LDA, we extracted new latent keywords that complemented the input provided by the user.

From now on, we will refer to the input keywords provided by the query as *input\_u*. Those latent keywords extracted by LDA will be referred to as *topic\_terms*.

To be noted, *topic\_terms set* provides information about the latent structure of the corpus and is independent of the query. *input\_u set* gives information about the query so that their main goal is to facilitate which terms and concepts from the corpus are desired by the user.

In order to obtain the *topic\_terms set*, we have processed each instance of the dataset in three phases: Pseudo-document generation, Text preprocessing and Topic modelling. Figure 1 shows a schematic representation of our model.

**3.3.1. Pseudo-document generation.** In this initial step, for every paper in the dataset, we aggregate it into a pseudo-document using the title of the paper, the text of its abstract and the titles of its references. The generation of pseudo-documents is a commonly used strategy to combat data sparsity [32].

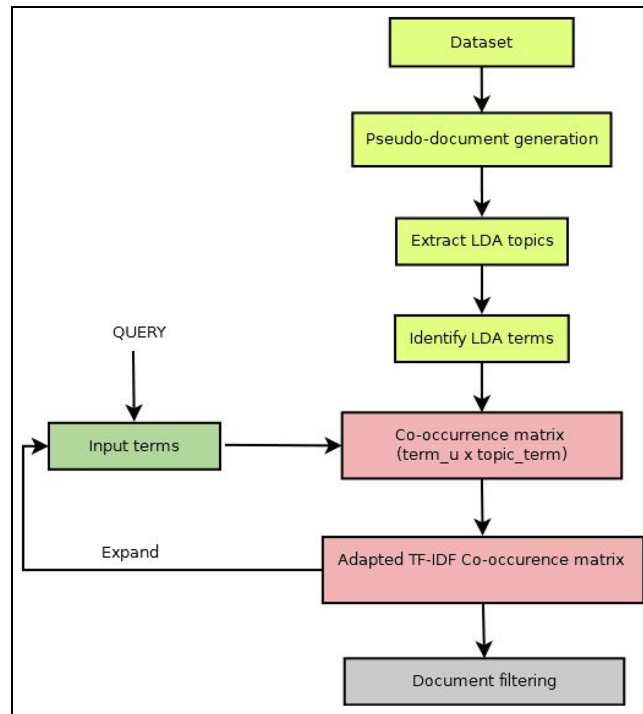
The use of the title and the abstract when extracting key information from a paper is quite common since it contains many keywords in a very limited space. The text of the paper is not usually taken into account since its content is usually quite heterogeneous, and the results vary greatly depending on the section being analysed [33]. That is why we have aggregated the title and the abstract and not the full text into the pseudo-document.

References are also a useful source of information in articles. Bibliographical coupling and co-citation were used in the very first approach to study relationships between articles [34]. In addition, references have been used to extract key concepts of a document in order to generate better keywords [35]. In the context of this work, where keywords are absent in the dataset, it is important to extract key concepts in order to help us to analyse the latent information of each article.

**3.3.1.1. Text preprocessing.** In this phase, we conducted a series of tasks that aim to transform the text into a more digestible form so that the irrelevant information is reduced and the LDA model can perform better.

To preprocess a pseudo-document, we performed the following tasks:

- The text is converted to lowercase, and the punctuation is deleted.
- We applied a *stop words* filter: Stop words are words that do not contribute any meaning to the document, such as articles, pronouns and prepositions. Removing stop words avoids generating style indicator keywords that are not useful in the desired context. We have used the list of clinical stop words provided by Ganesan et al. [36].
- Each word is lemmatised: The lemmatisation process is a linguistic process by which the root of a word is determined. We have used the spaCy lemmatiser algorithm.<sup>4</sup>



**Figure 1.** Schematic representation of the method.

Yellow squares represent steps required to obtain latent keywords in the dataset. The grey square represents the step that combines the latent information from the corpus with the user input terms.

- For reasons of efficiency, only unigrams are used in our model, but some n-grams have been kept for relevant reasons, those n-grams are as follows: *WHO, Public Health, Social Media, Fake News* and *Social Sciences*.
- Finally, papers written in languages other than English are eliminated in order to reduce the computing time in the keyword topic modelling step.

**3.3.1.2. Topic modelling.** LDA analysis was applied for each article of the dataset in order to extract the latent topics hidden in the documents. It is important to note that we have not applied the same model for the whole corpus. A new LDA model was generated for each article. Our goal here was to detect minority topics in the dataset that often can be hidden in a single article. Excess noise from popular topics could be introduced if the same LDA model was applied to the entire corpus.

After completing the above steps, we extracted the latent topics from the corpus. These latent topics do not depend on the user input but rather the documents inside the dataset. In any case, our goal here was not to use the topics extracted by LDA, but to work with the terms inside these topics. So, we combined the most relevant terms from each topic into a set of latent keywords. This set is called *LDA terms*.

The following steps combined the latent structure of the corpus with the concepts provided by the query.

**3.3.1.3. Co-occurrence matrix.** After modelling *LDA terms* for each article, we computed a co-occurrence matrix between all the input terms ( $term_u$ ) and all the LDA terms ( $term_{topic}$ ) for the whole corpus. So that we could count the number of times that a  $term_u$  co-occurs with a  $term_{topic}$ , in the corpus.

**3.3.2. Adapted TF-IDF co-occurrence matrix.** TF-IDF combines two metrics used in text information processing techniques: *term frequency* (TF) and *inverse document frequency* (IDF). TF is a metric used to represent the number of occurrences of a term in a document, while IDF indicates the number of times a term appears in a corpus. TF-IDF represents the importance of a term in the document. Considering important a term which appears frequently in a document but is rare in the corpus. This term can be used to represent the key information of the document.

In our study, we translated the idea behind TF-IDF and used it to analyse the relation between  $term\_topic$  and  $term\_u$ . TF-IDF's goal is to link the information provided by topic terms and input terms. The TF-IDF score obtained by each of the topic terms would depend on the query and, therefore, on the input terms.

First, we computed the topic frequency TF as the number of occurrences of an LDA term in all the LDA topics for each paper

$$TF(term) = \sum_{i=1}^p S_i(term) \quad (1)$$

where  $S_i(term) = 1$  if the term is in LDA terms  $\{term\_topic\}_i$  for the document  $i$ , and  $p$  is the total number of documents in the corpus.

Then, we computed the topic inverse document frequency (TIDF) as the inverse frequency of the number of times that a term occurs in the list of term topics for each article

$$TIDF(term) = \log\left(1 + \frac{p}{TF(term)^c}\right) \quad (2)$$

where  $c$  is a constant number. A greater value of  $c$  would indicate a lower TIDF score in relation to frequent terms, while smaller  $c$  values would indicate a better TIDF score in relation to frequent terms. For our dataset, we have used  $c = 2$  as the value that allowed us to extract latent terms with similar popularity to the user inputs. Readers can refer to the Supplemental material section to see further analysis of the behaviour of the  $c$  value. A deeper justification about the use of the adapted version of TF-IDF that we have proposed in our article will be further explained in the section *Energy and c parameter behaviour*.

We defined LDA-Term Co-occurrence (LTC) as the number of times that a  $term\_u$  co-occurs with a  $term\_topic$ . This value is provided by the co-occurrence matrix defined above

$$LTC(term\_topic, term\_u) = C(term\_topic, term\_u) \quad (3)$$

where  $C$  is the co-occurrence matrix.

Our adapted TF-IDF metric is expressed as follows

$$ATF.IDF(term\_topic, term\_u) = LTC(term\_topic, term\_u) \cdot TIDF(term\_topic) \quad (4)$$

In our method, LTC fulfils the same role as term frequency in TF-IDF [37]. We worked with term topics instead of documents, so the aim of LTC is to weigh terms inside of term topics. The more frequent a term is present inside a term topic, the more important this term becomes for that term topic. LTC is a simple way to measure the importance of a term within a term topic. In the same way, if a term is very frequent in each term topic, then this term is not relevant or may give us irrelevant information. Therefore, the objective of TIDF is to present a high score to less frequent terms. Finally, we combined LTC with TIDF results to get a high score from those very frequent terms in only a few terms topics which are, at the same time, very infrequently seen throughout the rest of the term topics.

Finally, we built an adapted TF-IDF co-occurrence Matrix (ATF.IDF)

$$ACO(term\_topic, term\_u) = ATF.IDF(term\_u, term\_topic) \quad \forall term\_u \in U, \forall term\_topic \in T \quad (5)$$

where  $T$  is a set with all LDA term topics found in the corpus and  $U$  are the input terms.

**3.3.2.1. Keywords expansion.** In this step, we expanded the input terms with new topics extracted from LDA terms. From the ATF.IDF co-occurrence matrix built in the previous step, we computed the sum of the ATF.IDF scores for all the LDA terms topic as

$$total\_score(ACO) = \sum_{term\_u \in U} \sum_{term\_topic \in T} ACO(term\_topic, term\_u) \quad (6)$$

Then, we applied an energy threshold to the score. This energy threshold is a classical mechanism to preserve information until reaching a specific threshold score [38]

$$threshold(ACO, energy) = total\_score(ACO) \cdot energy \quad (7)$$

Term topics are ranked according to their ATF-IDF score so that the term topics with greater scores appear first

$$score(term_{topic}) = \sum_{term_u \in U} ACO(term_{topic}, term_u) \quad (8)$$

The next step computed the cumulative sum of all term topics, according to their score, and selected only the LDA terms whose cumulative ATF-IDF score was under the threshold. Finally, these selected LDA terms were added to the input terms set

$$selectedTerms(ACO, energy) = \{t_1, t_2, \dots, t_n\} \{t_j \in LDA \wedge cumsum(score(t_j)) < threshold(ACO, energy)\} \quad (9)$$

We can thus repeat all the steps again using the new input terms in order to find new latent keywords. An example of keywords expansion can be seen in Table 3.

**3.3.3. Document filtering.** For this final step, documents were filtered according to the expanded input terms set ( $\{term_u\}$ ). So the filters were built using the initial user input terms, and the expanded input terms derived from the latent structure. Next, we extracted all the documents from the corpus that contained the keywords. The minimum number of keywords that had to be in a document can be set as an additional threshold. In our case, we have imposed a minimum of two keywords.

### 3.4. Time complexity

The time complexity was mainly affected by the execution of the LDA method and is  $O(p \cdot (mnt + t^3))$  where  $p$  is the number of LDA documents in the corpus,  $m$  is the number of topics,  $n$  is the number of terms, and  $t = \min(m, n)$  [39]. While this topic is not directly related to our work, there are some proposals to reduce the time complexity of LDA algorithms [39,40] that can be useful for readers.

### 3.5. Experimental set-up

In this article, we have executed the proposed method in three iterations, in order to extract three levels of keywords. After performing a hyperparameter optimisation search using the Grid Search method provided by the Scikit-learn library,<sup>5</sup> the best energy threshold for our dataset was 0.025.

The  $c$  parameter of TIDF was 2 (Please refer to Section 4 and check the Supplemental material of this article for more details about ATF.IDF).

In respect to LDA, the number of topics was set to 30, which gave us good results in terms of coherence. We used the UMass-Coherence to determine the number of topics [41].

## 4. Results

In this section, we will proceed to show the results obtained using our proposed method. The goal of our Information Retrieval System was to find an answer to the question: ‘What has been published about ethical and social science considerations?’. In addition, specifically, we wanted information from seven different thematic areas (See Table 1).<sup>6</sup>

As explained in previous sections, the dataset provided by Kaggle does not contain information about the keywords of the articles, so we could only work with the title, abstract and references of each article. The absence of keywords is a challenge when identifying the category in which an article is framed, since the title and abstract do not always contain enough information on all the topics covered by a scientific work.

In this context, evaluating the performance of our method can be a challenging problem. Fortunately, the provided dataset contains enough information to identify the articles that compose the corpus; thus, we were able to perform a manual evaluation in two steps:

1. *A priori evaluation:* By checking the titles and abstracts of the selected article, it was verified whether the articles selected by our method could answer the question posed. This evaluation was performed exclusively using the data provided by the dataset.

**Table 1.** Thematic areas.

Numbers	Thematic areas
1	Efforts to articulate and translate existing ethical principles and standards to salient issues in COVID-19.
2	Efforts to embed ethics across all thematic areas, engage with novel ethical issues that arise and coordinate to minimise duplication of oversight.
3	Efforts to support sustained education, access, and capacity building in the area of ethics.
4	Efforts to establish a team at World Health Organization that will be integrated within multidisciplinary research and operational platforms and that will connect with existing and expanded global networks of social sciences.
5	Efforts to develop qualitative assessment frameworks to systematically collect information related to local barriers and enablers for the uptake and adherence to public health measures for prevention and control. This includes the rapid identification of the secondary impacts of these measures. (e.g. use of surgical masks, modification of health-seeking behaviours for SRH, school closures).
6	Efforts to identify how the burden of responding to the outbreak and implementing public health measures affects the physical and psychological health of those providing care for COVID-19 patients and identify the immediate needs that must be addressed.
7	Efforts to identify the underlying drivers of fear, anxiety and stigma that fuel misinformation and rumour, particularly through social media.

**Table 2.** Thematic areas and initial input terms.

Thematic areas	Initial input terms
1	ethic, moral, fair, justice, immoral, standard
2	ethic, oversight, justice, care, sociology, education, anthropology, bibliometric, social, moral, dilemma, concerns
3	education, access, ethics, fellowship, teaching, principles, philosophy, students, training
4	World Health Organization, research, global, multidisciplinary, social, science, university, collaboration
5	local, barrier, public, measures, society, pandemic, enablers, publicHealth, prevention, control, impact, closures, quarantine
6	outbreak, publicHealth, public, measures, psychology, care, COVID-19, needs, urgently, response, resiliency, pandemic, nurse, medic, employee, professional, worker
7	stigma, misinformation, rumor, socialMedia, media, news, papers, networks, fake, fakeNews, facebook, twitter

2. *A posteriori evaluation*: By manually obtaining each one of the selected articles, the articles were checked to see whether they would be able to answer the question posed. This evaluation was performed by using external data to the provided corpus by the dataset.

In addition, keywords of the selected articles were extracted and compared with the topics generated with our method in order to check the level of co-occurrence. Figure 2 shows an example of co-occurrence matrix generation.

For each thematic area, a list of keywords was chosen, as shown in Table 2. These keywords are the initial input terms that compound our queries and were chosen manually.

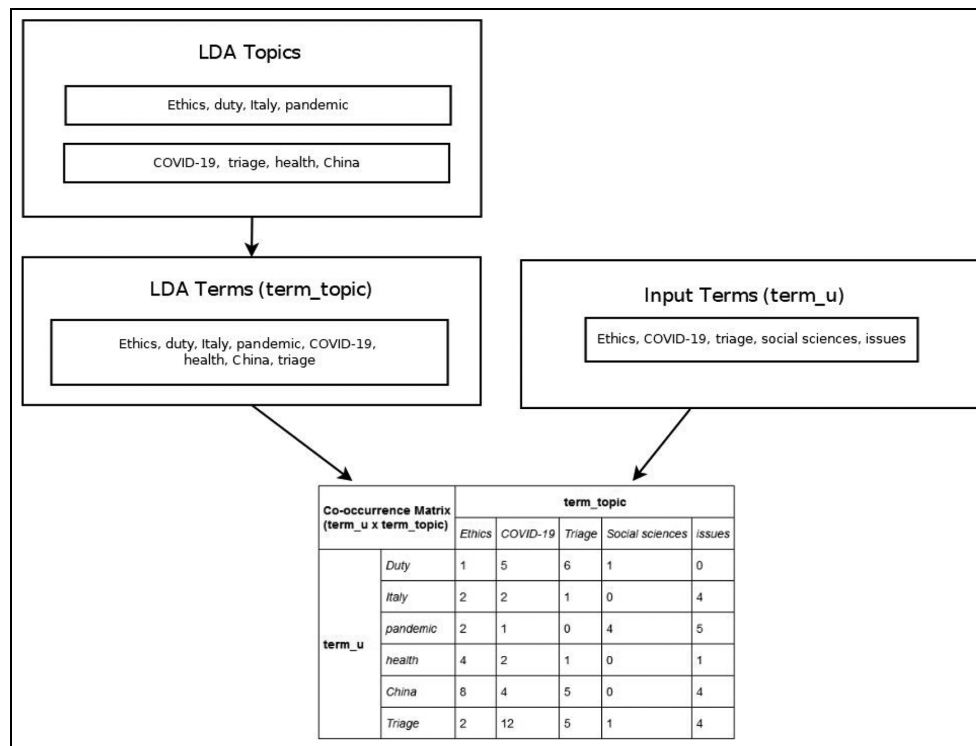
#### 4.1. *A priori evaluation*

During ‘a priori’ evaluation, the top 10 most relevant papers for each thematic area were selected. We performed two surveys with the goal of checking whether the filtered papers were able to answer the question posed in the corresponding thematic area. To this end, in the first survey, the participants had to evaluate whether the title and the abstract of a set of articles were related to a thematic area. Thus, participants were only allowed to see the same information used by the method to filter and retrieve the selected papers. Participants were randomly assigned to two thematic areas and had to read the title and abstract of ten articles retrieved by our method for the assigned thematic areas.

There were three possible answers:

1. The article fits the thematic area.
2. The article does not fit the thematic area.
3. With the provided information, we cannot know whether the article fits the thematic area or not.





**Figure 2.** Co-occurrence matrix generation.

**Table 3.** Example of filtering documents and input terms expansion after three expansions.

Expansion (#iteration)	Input terms	Documents
1	Ethic, Moral, Fair, Justice, Immoral, Standard	260
2	Ethic, Moral, Fair, Justice, Immoral, Standard, <b>Consent, Bioethic, Duti, Principi</b>	306
3	Ethic, Moral, Fair, Justice, Immoral, Standard, Consent, Bioethic, Duti, Principi, <b>Oblig, Alloc</b>	319

Bold terms are introduced from LDA terms during the previous expansion. Document column indicates the number of filtered documents. Note that terms are stemmed.

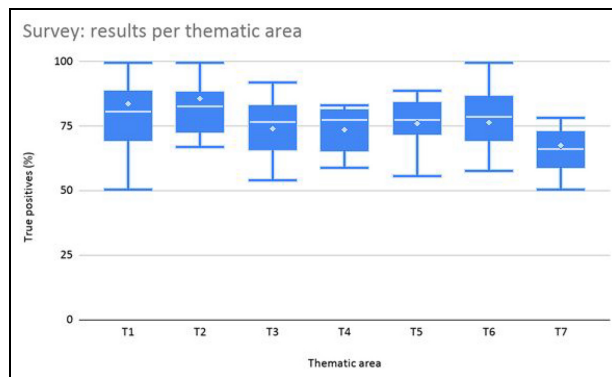
Answers 1 and 2 helped to evaluate an article as a true positive and a false positive, respectively.

Table 4 shows the results obtained in the survey per thematic area. The range of different responses given by the participants can be seen in Figure 3.

The survey was answered by 57 participants, all of them use Amazon Mechanical Turk,<sup>7</sup> a web platform where users get an economic reward for doing tasks that require human intelligence. All participants were native English speakers. The time to complete the survey and response patterns were analysed in order to prevent random responses.

Table 4 shows the responses obtained in the survey. For all the thematic areas, the first answer (The article fits the thematic area) was the most common, being selected for the majority of articles. According to our results, the second thematic area ('Efforts to embed ethics across all thematic areas, engage with novel ethical issues that arise, and coordinate to minimize duplication of oversight'.) obtained the best results, with about 81% of users answering with the first answer while the last thematic area ('Efforts to identify the underlying drivers of fear, anxiety, and stigma that fuel misinformation and rumors, particularly through social media'.) achieved the worst results, with a score of only 67.144% for the first answer.

Figure 3 shows the answer interval for every thematic. That means that, for example, in the first thematic area there was an article that was classified as a true positive by 50% of the participants. At the same time, there was an article that was classified as a true positive by 100% of the participants assigned to that thematic area. In addition, in the



**Figure 3.** Results per thematic area. We can see the range of true positive values obtained by our method according to the different responses given by the participants in the survey.

**Table 4.** Responses per thematic area (See Table 1 for information about each thematic area).

Thematic areas	Response 1 (%)		Response 2 (%)		Response 3 (%)	
	Mean	SD	Mean	SD	Mean	SD
1	79	15.23	16	11.73	4	5.16
2	81.03	11.79	13.42	10.29	5.56	5.85
3	73.59	11.72	19.42	12.24	6.15	4.86
4	73.031	10.96	21.113	9.20	5.832	5.62
5	76.67	9.72	15.55	10.73	7.77	5.36
6	77.141	13.08	15.717	12.51	7.14	12.14
7	67.144	9.03	21.428	9.52	11.429	4.99

SD: standard deviation.

Response 1 = The article fit the thematic area; 2 = The article does not fit the thematic area; and 3 = With the provided information, it is not possible to determine whether the article fits the thematic area or not.

**Table 5.** Confusion matrix representing results from the second survey.

Thematic areas	1	2	3	4	5	6	7
1	8	1	0	1	0	1	0
2	2	8	1	0	2	0	0
3	0	1	7	1	1	0	0
4	0	0	1	8	2	1	1
5	0	0	0	0	5	0	0
6	0	0	0	0	0	5	3
7	0	0	1	0	0	3	6

Rows represent the participants' responses, while columns represent the estimated classification performed by our method. Highlighted cells represent True positives.

Supplemental material section, we have included a Figure where readers can get additional information about the precision obtained in the survey results.

In the second survey, we randomly divided the same papers from the first survey into five groups, ensuring that each group had two papers per subject area (14 papers per group). Then, we arbitrarily assigned each participant to a single group.

Participants then had to read the abstract and the title of each paper in their assigned group and assign a thematic area. They also had the option of not assigning any thematic area if they felt unable to classify the paper.

**Table 6.** Metrics extracted from the second survey.

TA	TP	FP	FN	Precision	Recall	TN	TNR	Accuracy	F-score
1	8	2	3	0.73	0.8	39	0.95	0.90	0.76
2	8	2	5	0.62	0.8	39	0.90	0.87	0.70
3	7	3	3	0.70	0.7	40	0.93	0.89	0.70
4	8	2	5	0.62	0.8	39	0.95	0.87	0.70
5	5	5	0	1.00	0.5	42	0.89	0.90	0.67
6	5	5	3	0.63	0.5	42	0.89	0.85	0.56
7	6	4	4	0.60	0.6	41	0.91	0.85	0.60

TA: thematic area; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; TNR: true/negative rate.

**Table 7.** Confusion matrix presenting results from 'a posteriori' results.

Thematic areas	1	2	3	4	5	6	7
1	7	1	3	1	2	0	0
2	2	8	1	0	0	0	0
3	0	1	6	1	0	1	0
4	1	0	0	8	0	0	0
5	0	0	0	0	8	0	0
6	0	0	0	0	0	7	2
7	0	0	0	0	0	2	8

Rows represent the real classification, while columns represent the estimated classification performed by our method. Highlighted cells represent True positives.

For the second survey, we had 65 participants from Amazon Mechanical Turk and all participants were native English speakers. As in the previous survey, we analysed the time and response patterns to prevent random responses. Results from the second survey are presented in the confusion matrix shown in Table 5. From this confusion matrix, some metrics have been extracted and can be found in Table 6. As we can see, precision is, again, obtaining high scores. At this point, it is important to state that the second survey required participants to perform a harder task than in the first survey, so lower scores were expected. The recall also had high scores except for the third thematic area, where results were  $< 0.5$ . Also, accuracy and  $F$ -scores were over 0.5 points for each thematic area.

Precision, recall, accuracy, true/negative rates and  $F$ -scores were calculated by using definitions from Powers [42].

Despite the results obtained, 'a priori' evaluation has several limitations. First, it is still difficult, even for humans, to decide what the content of an article is by only checking its title and abstract. Second, while there are documents with very long abstracts, other articles have short or even non-existent abstracts. In addition, the dataset did not exclusively contain scientific papers, and it was possible to encounter editorial articles and other documents whose summaries were not homologated with abstracts.

In any case, 'a priori' analysis can be very useful in evaluating results obtained when working with the same conditions as used in our method.

## 4.2. A posteriori evaluation

To perform the 'a posteriori' evaluation, we downloaded and manually classified 150 papers by using the following guidelines:

1. Check whether the title, abstract and references contained terms or topics related to the thematic area.
2. Check whether keywords matched topics from the thematic area.
3. Check whether conclusions contained keywords or topics related to the thematic area.
4. Read the full article if the previous steps did not provide enough evidence to make a decision.

After this manual classification, from the subset of 150 previously classified papers, we randomly extracted 10 papers per thematic area and tested the performance of our model in terms of precision, accuracy, recall and  $F$ -scores. This type of evaluation allowed us to more precisely know whether an article was appropriately selected because we could access

**Table 8.** Metrics extracted from the ‘a posteriori’ evaluation.

TA	TP	FP	FN	Precision	Recall	TN	TNR	Accuracy	F-score
1	7	3	7	0.50	0.7	45	0.94	0.84	0.58
2	8	2	3	0.73	0.8	44	0.96	0.91	0.76
3	6	4	3	0.67	0.6	46	0.92	0.88	0.63
4	8	2	1	0.89	0.8	44	0.96	0.94	0.84
5	8	2	0	1.00	0.8	44	0.96	0.96	0.89
6	7	3	2	0.78	0.7	45	0.94	0.91	0.74
7	8	2	2	0.80	0.8	44	0.96	0.93	0.80

TA: thematic area; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; TNR: true/negative rate.

information from the whole document (title, authors, keywords, abstract, references and text). ‘A posteriori’ evaluation was then performed by the authors. ‘A posteriori’ results are described in Tables 7 and 8.

If we compare both evaluation strategies, ‘a priori’ (second survey) and ‘a posteriori’, as noted, the ‘a posteriori’ evaluation gets better scores for nearly all thematic areas. In T5, both evaluations scored a 100% in terms of precision.

The lack of information during the ‘a priori’ evaluation (first survey) was clearly reflected during the second step of the evaluation process. For example, for the first thematic area, one of the articles retrieved by the method was titled: ‘Understanding perceptions of global healthcare experiences on provider values and practices in the USA: a qualitative study among global health physicians and program directors’ [43].

The title and the abstract of the article did not contain words explicitly related to the thematic area like ‘ethics’, ‘principles’ and ‘standards’ so that it was more difficult to evaluate whether the article fit the topic or not. In addition, the article was not about COVID-19 disease or any other pandemic. Nevertheless, by carefully checking the article, it was clear that the article was a true positive. This article achieved the worst results in its thematic area, where only 50% of the participants believed that the article fit the topic, while 40% of participants believed that the article did not fit the topic. Meanwhile, our method retrieved this article as one of the top 10 for the thematic area.

From these results, we can say that it seems that using certain methods to extract the latent structure of a document can help us establish relationships between words that are not necessarily evident to the human mind. Though this is not the scope of our work, further analysis should be done.

On the contrary, there are two thematic areas where ‘a priori’ evaluation overcame ‘a posteriori’ results. For example, in T4, our method retrieved an article titled ‘The Highest Cited Papers on Brucellosis: Identification Using Two Databases and Review of the Papers’ Major Findings’ [44]. This article is a bibliometric analysis but is not connected with key concepts to this thematic area, like ‘Web Health Organization’ or global network of social sciences, and yet most participants answered that the article fit the topic.

### 4.3. Energy and $c$ parameter behaviour

As mentioned before, energy and  $c$  were parameters used in equations (7), (9) (energy), and equation 2 ( $c$ ). We performed an ‘a posteriori’ evaluation strategy to study the behaviour of energy and the  $c$  parameter. According to our tests, the best performance was achieved when  $c$  was 2 and energy was 0.02. Nevertheless, it is important to note that every thematic area had a different score. For example, for thematic areas 2 and 3,  $c = 3$  gave a better performance, while energy = 0.25 was the best option for thematic area 2. Figures 4 and 5 illustrate the behaviour of  $c$  and energy parameters for each thematic area, while Figure 6 shows an average  $F$ -score across all thematic areas.

### 4.4. The role of ATF.IDF

In Section 3, we introduced ATF.IDF, and we adapted the TF-IDF proposal. Figure 7 shows a comparison of the keyword expansion stage according to the value of  $c$  in terms of *popularity*. While Figure 7 groups terms by their stage in the method, Figure 8 shows the same information but clusters terms by popularity. The popularity of a term is the number of articles in the corpus where the term appears. For example, if term A appears in 2500 articles and term B appears in 12 documents, the popularity of term A (popularity = 2500) is greater than the popularity of term B (popularity = 12). At the same time, we can say that term B is a more specific term than term A.

In both figures, we analysed the input terms of thematic area number 5 (See Table 2) If we use  $c = 1$ , the input terms are expanded with the LDA topics described below:

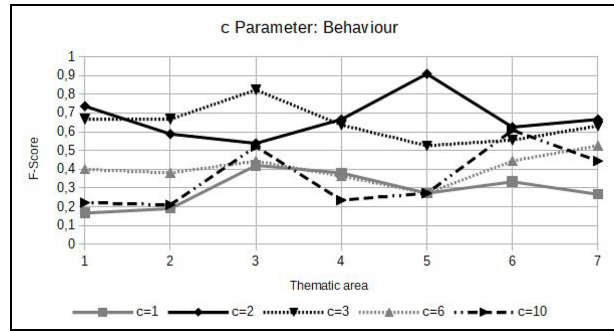


Figure 4. Behaviour of the c parameter. Energy is fixed at 0.025.

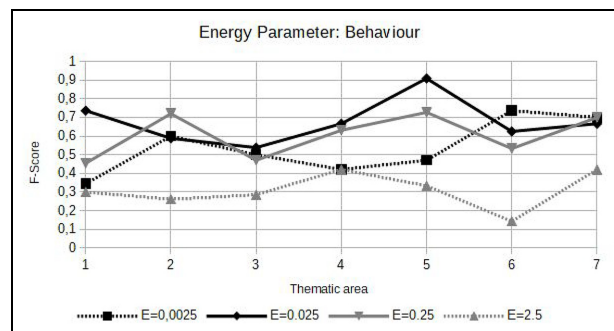


Figure 5. Behaviour of the energy parameter c is fixed at 2.

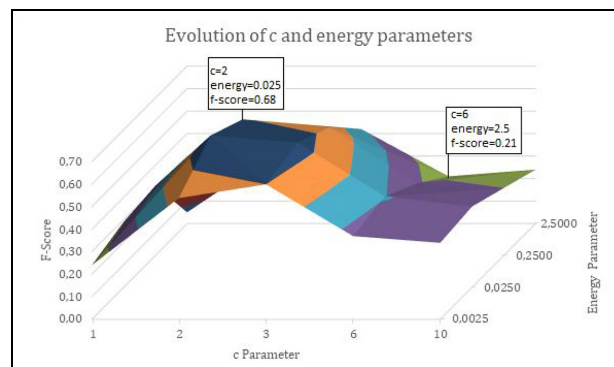


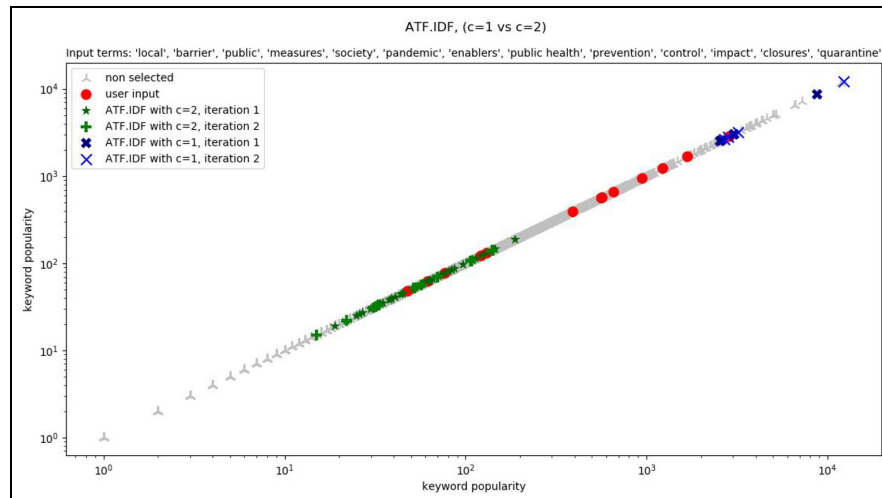
Figure 6. Surface plot representing F-score evolution according to c and energy parameters. The F-score is measured as the mean F-score for whole thematic areas.

- First iteration: ‘health’ and ‘diseas’, ‘influenza’, ‘outbreak’.
- Second iteration: ‘viru’, ‘infect’, ‘respiratori’, ‘epiderm’.

Finally, in Figure 8, elements represented with a grey mark are terms that had relations with the query but were not selected.

When using  $c = 2$ , expanded terms were represented by a diamond in Figure 7 for terms extracted during the first iteration and with a star for the second iteration.





**Figure 8.** Comparison of ATF, IDF selected items according to  $c$  value.  $c = 1$  versus  $c = 2$ . Thematic area #5. Sorted by popularity.

Our method can be a valuable tool when used as an Information Retrieval System, with the capacity to focus on retrieving specific information about complex topics in a dataset where the thematic area has a secondary role. Nevertheless, we have performed our analysis working with the corpus provided by the White House in collaboration with the Allen Institute, where only abstracts and titles were presented in the corpus so that the full semantic information from papers has not been analysed. This limitation should be taken into consideration.

For future works, it is necessary to analyse the behaviour of this method in different thematic areas and to test our work with a full semantic information corpus.

### Declaration of conflicting interests


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

### Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

### ORCID iDs

Jorge Chamorro-Padial  <https://orcid.org/0000-0002-6334-3786>

Francisco-Javier Rodrigo-Ginés  <https://orcid.org/0000-0001-6235-6860>

### Notes

1. <https://www.research.ibm.com/covid19/deep-search/> (accessed 28 December 2021).
2. <https://www.semanticscholar.org/cord19>.
3. <https://scikit-learn.org/stable/index.html> (accessed 3 June 2020).
4. [https://www.nltk.org/\\_modules/nltk/stem/wordnet.html](https://www.nltk.org/_modules/nltk/stem/wordnet.html) (accessed 31 January 2022).
5. <https://scikit-learn.org/stable/> (accessed 31 January 2022).
6. The questions and thematic areas referred in our work were extracted from the Kaggle Challenge.
7. <https://www.mturk.com/> (accessed 3 June 2020).

### Supplemental material

Supplemental material for this article is available online.

### References

- [1] Hui DS, Azhar EI, Madani TA et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health – the latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis* 2020; 91: 264–266.

- [2] World Health Organization (WHO). WHO Director-General's opening remarks at the media briefing on COVID-19 – 6 March 2020, <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-6-march-2020> (2020, accessed 26 May 2020).
- [3] Chahrour M, Assi S, Bejjani M et al. A bibliometric analysis of COVID-19 research activity: a call for increased output. *Cureus* 2020; 12: e7357.
- [4] Atkeson AG. *What will be the economic impact of COVID-19 in the US? Rough estimates of disease scenarios*. Los Angeles, 2020, <https://www.minneapolisfed.org/research/staff-reports/what-will-be-the-economic-impact-of-covid-19-in-the-us-rough-estimates-of-disease-scenarios>
- [5] Li S, Wang Y, Xue J et al. The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. *Int J Environ Res Public Health* 2020; 17: 2032.
- [6] Liu Q, Zheng Z, Zheng J et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. *J Med Internet Res* 2020; 22: e19118.
- [7] Huynh TLD. The COVID-19 risk perception: a survey on socioeconomics and media attention. *Econ Bullet* 2020; 40: 758–764.
- [8] Torres-Salinas D. Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en Bases de Datos y Repositorios en Acceso Abierto. *Prof Inform*. Epub ahead of print 14 April 2020. DOI: 10.3145/epi.2020.mar.15.
- [9] COVID-19 open research dataset challenge (CORD-19) Kaggle, <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> (accessed 30 January 2022).
- [10] Lou J, Tian SJ, Niu SM et al. Coronavirus disease 2019: a bibliometric analysis and review. *Eur Rev Med Pharmacol Sci* 2020; 24: 3411–3421.
- [11] Nasab FR and Rahim F. Bibliometric analysis of global scientific research on SARSCoV-2 (COVID-19). *medRxiv*. Epub ahead of print 23 March 2020. DOI: 10.1101/2020.03.19.20038752.
- [12] Bonnevie E, Gallegos-Jeffrey A, Goldburg J et al. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *J Comm Healthc* 2021; 14: 12–19.
- [13] Baraybar-Fernández A, Arrufat-Martín S and Rubira-García R. Public information, traditional media and social networks during the COVID-19 crisis in Spain. *Sustainability* 2021; 13: 6534.
- [14] Feng Y and Zhou W. Is working from home the new norm? *An observational study based on a large geo-tagged COVID-19 Twitter dataset*, <http://arxiv.org/abs/2006.08581> (2020, accessed 28 December 2021).
- [15] Sharifi A and Khavarian-Garmsir AR. The COVID-19 pandemic: impacts on cities and major lessons for urban planning, design, and management. *Sci Total Environ* 2020; 749: 142391.
- [16] Cinelli M, Quattrocioni W, Galeazzi A et al. The COVID-19 social media infodemic, <http://arxiv.org/abs/2003.05004> (2020, accessed 26 May 2020).
- [17] Lopez CE, Vasu M and Gallemore C. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset, <http://arxiv.org/abs/2003.10359> (2020, accessed 26 May 2020).
- [18] Singh L, Bansal S, Bode L et al. A first look at COVID-19 information and misinformation sharing on Twitter, <http://arxiv.org/abs/2003.13907> (2020, accessed 26 May 2020).
- [19] Schild L, Ling C, Blackburn J et al. 'Go eat a bat, Chang!': on the emergence of Sinophobic behavior on web communities in the face of COVID-19, <http://arxiv.org/abs/2004.04046> (2020, accessed 26 May 2020).
- [20] Riloff E, Schafer C and Yarowsky D. Inducing information extraction systems for new languages via cross-language projection. In: *Proceedings of the 19th international conference on computational linguistics*, Taipei, Taiwan, 24 August–1 September 2002, pp. 1–7. Stroudsburg, PA: Association for Computational Linguistics (ACL).
- [21] Latif S, Usman M, Manzoor S et al. Leveraging data science to combat COVID-19: a comprehensive review. *IEEE T Artif Intel* 2020; 1: 85–103.
- [22] Shah PK, Perez-Iratxeta C, Bork P et al. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics* 2003; 4: 20.
- [23] Williams R. *Keywords: a vocabulary of culture and society*. 2nd ed. New York: Oxford University Press, 1985.
- [24] Baker P. Querying keywords: questions of difference, frequency, and sense in keywords analysis. *J Engl Linguist* 2004; 32: 346–359.
- [25] Wang X, Liu W, Chauhan A et al. Automatic textual evidence mining in COVID-19 literature, <http://arxiv.org/abs/2004.12563> (2020, accessed 28 December 2021).
- [26] Esteva A, Kale A, Paulus R et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj Digit Med* 2021; 4: 1–9.
- [27] Voorhees E, Alam T, Bedrick S et al. TREC-COVID: constructing a pandemic information retrieval test collection. *ACM SIGIR Forum* 2020; 54: 1.
- [28] Best P, Taylor B, Manktelow R et al. Systematically retrieving research in the digital age: case study on the topic of social networking sites and young people's mental health. *J Inf Sci* 2014; 40: 346–356.
- [29] Karlsson A, Hammarfelt B, Steinhauer HJ et al. Modeling uncertainty in bibliometrics and information retrieval: an information fusion approach. *Scientometrics* 2015; 102: 2255–2274.
- [30] Dimitrakis E, Sgontzos K and Tzitzikas Y. A survey on question answering systems over linked data and documents. *J Intell Inf Syst* 2020; 55: 233–259.



- [31] Mohammed SH and Al-Augby S. LSA & LDA topic modeling classification: comparison study on E-books. *Indones J Electr Eng Comput Sci* 2020; 19: 353–362.
- [32] Zuo Y, Wu J, Zhang H et al. Topic modeling of short texts: a pseudo-document view. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, CA, 13–17 August 2016, pp. 2105–2114. New York: Association for Computing Machinery (ACM).
- [33] Chen G and Xiao L. Selecting publication keywords for domain analysis in bibliometrics: a comparison of three methods. *J Informetr* 2016; 10: 212–223.
- [34] Weinberg BH. Bibliographic coupling: a review. *Inform Storage Ret* 1974; 10: 189–196.
- [35] Garfield E. KeyWords Plus – ISI’s breakthrough retrieval method. 1. Expanding your searching power on current-contents on diskette. *Curr Contents* 1990; 1: 5–9.
- [36] Ganesan K, Lloyd S and Sarkar V. Discovering related clinical concepts using large amounts of clinical notes. *Biomed Eng Comput Biol* 2016; 7(suppl. 2): 27–33.
- [37] Roelleke T and Wang J. TF-IDF uncovered: a study of theories and probabilities. In: *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (ACM SIGIR’2008)*, Singapore, 20–24 July 2008, pp. 435–442. New York: ACM Press.
- [38] Singla A and Patra S. A fast automatic optimal threshold selection technique for image segmentation. *Signal Image Video P* 2017; 11: 243–250.
- [39] Cai D, He X and Han J. Training linear discriminant analysis in linear time. In: *Proceedings of the 2008 IEEE 24th international conference on data engineering*, Cancun, Mexico, 7–12 April 2008, pp. 209–217. New York: IEEE.
- [40] Sontag D and Roy DM. Complexity of inference in latent Dirichlet allocation, 2011, <https://papers.nips.cc/paper/2011/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html>
- [41] Mimno D, Wallach H, Talley E et al. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, Edinburgh, 27–31 July 2011, pp. 262–272. Stroudsburg, PA: Association for Computational Linguistics (ACL).
- [42] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, <http://arxiv.org/abs/2010.16061> (2020, accessed 30 January 2022).
- [43] Matthews-Trigg N, Citrin D, Halliday S et al. Understanding perceptions of global healthcare experiences on provider values and practices in the USA: a qualitative study among global health physicians and program directors. *BMJ Open* 2019; 9: e026020.
- [44] Bakri FG, Alqadiri HM and Adwan MH. The highest cited papers in brucellosis: identification using two databases and review of the papers’ major findings. *Biomed Res Int* 2018; 2018: 9291326.