PLOS ONE

RESEARCH ARTICLE

An active learning-based approach for screening scholarly articles about the origins of SARS-CoV-2

Xin An[1], **Mengmeng Zhang**[1], **Shuo Xu**[2]*

1 School of Economics & Management, Beijing Forestry University, Beijing, P.R. China, **2** College of Economics and Management, Beijing University of Technology, Beijing, P.R. China

* xushuo@bjut.edu.cn

Citation: An X, Zhang M, Xu S (2022) An active learning-based approach for screening scholarly articles about the origins of SARS-CoV-2. PLoS ONE 17(9): e0273725. https://doi.org/10.1371/journal.pone.0273725

Editor: Zahid Mehmood, University of Engineering and Technology Taxila Pakistan, PAKISTAN

Received: December 8, 2021

Accepted: August 13, 2022

Published: September 16, 2022

Data Availability Statement: All data files are available at https://github.com/pzczxs/COVID-Origin.

Funding: X An received the National Natural Science Foundation of China under grant number 72004012, and S Xu received the National Natural Science Foundation of China under grant number 72074014 and Special Exploration Project for Anti-epidemic by Beijing University of Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

To build a full picture of previous studies on the origins of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), this paper exploits an active learning-based approach to screen scholarly articles about the origins of SARS-CoV-2 from many scientific publications. In more detail, six seed articles were utilized to manually curate 170 relevant articles and 300 nonrelevant articles. Then, an active learning-based approach with three query strategies and three base classifiers is trained to screen the articles about the origins of SARS-CoV-2. Extensive experimental results show that our active learning-based approach outperforms traditional counterparts, and the uncertain sampling query strategy performs best among the three strategies. By manually checking the top 1,000 articles of each base classifier, we ultimately screened 715 unique scholarly articles to create a publicly available peer-reviewed literature corpus, *COVID-Origin*. This indicates that our approach for screening articles about the origins of SARS-CoV-2 is feasible.

1. Introduction

In December 2019, a novel coronavirus SARS-CoV-2 caused a serious outbreak of acute respiratory disease [1]. This has brought the epidemic into the field of vision of human beings again, and the outbreak is still ongoing in many countries and territories. To completely block the spread of the epidemic and to further prevent similar or more serious epidemics in the future, the most fundamental task is to find the origins of SARS-CoV-2 and clarify how it reaches the human population [2]. In human history, the sources of many viruses are very difficult to trace [3,4]. For the purpose of successfully tracing the origins of a virus, multiple steps are involved in this procedure: epidemiological investigation, genome analysis, intermediate and natural host identification, field sampling, homology analysis of the virus strain and so on. To resolve this complicated puzzle, it is necessary for scientists to build a full picture of previous studies on the origins of SARS-CoV-2 and remain up to date on the latest ones.

In the literature, there is an explosive growth in scientific publications on COVID-19 (Corona Virus Disease 2019) and SARS-CoV-2 [5–7]. Zuo et al. [8] observed that 128 unique datasets on SARS-CoV-2 and COVID-19, including LitCovid [9] and CORD-19 [10], have

Competing interests: The authors have declared that no competing interests exist.

been developed and updated regularly for different tasks. However, to the best of our knowledge, there is no benchmark literature dataset publicly available for the origins of this virus. To bridge this gap and to promote global cooperation, this study is devoted to screening scholarly articles about the origins of SARS-CoV-2 from large numbers of scientific publications.

A naïve solution for screening articles is to carefully design a search strategy and then to retrieve the related publications from a comprehensive bibliographic database. Although this solution is very popular in practice, precision and recall are still the most concerning issues [11,12]. Take the strategy "TS = (SARS-CoV-2) AND TS = (origin)" in the Web of Science as an example. Several irrelevant articles, such as [13,14], appear in the result list since their titles or abstracts simultaneously contain the keywords *SARS-CoV-2* and *origin*. In addition, multiple relevant publications are missed, such as [15,16]. An alternative solution is to see document screening as a binary text classification problem. However, many text classification methods rely on the availability of a large labeled corpus. Due to an unprecedented volume of academic articles published related to this epidemic, it is not realistic to manually annotate enough samples for a text classification method with satisfactory performance. To meet this challenge, this study proposes an active learning-based approach for screening scholarly articles about the origins of SARS-CoV-2 with the following main contributions:

- An active learning-based approach is proposed to screen scientific publications about the origins of SARS-CoV-2.
- A curated peer-reviewed literature corpus (COVID-Origin), which can be freely accessed at https://github.com/pzczxs/COVID-Origin, was developed to track up-to-date peer-reviewed studies on the origins of SARS-CoV-2.
- Extensive experiments indicate that our approach, especially armed with multiple base classifiers, can efficiently screen scholarly articles about the origins of SARS-CoV-2.

The rest of the article is organized as follows. After briefly reviewing related work in Section 2, we describe the detailed process of data collection, annotation, and document representation in Section 3. Then, an active learning-based framework is developed in more detail in Section 4, and extensive experiments are conducted in Section 5. Section 6 concludes this contribution with the possible limitations of our study and future research.

2. Related work

2.1. Automatic document screening

As its name states, automatic document screening automatically finds all relevant documents pertinent to a given topic. Hence, this problem is also referred to as *the total recall problem* in the field of information retrieval [17]. More specifically, this problem can be formally described as follows. Given a set of candidate documents, of which only a small fraction is positive, each candidate can be checked to determine its label as positive or negative. The task is to check and label as few candidates as possible while achieving very high recall.

Since the work of Counsell [18], many approaches have been developed in the literature. It is well motivated in many applications, including systematic reviews in evidence-based medicine [19,20] and software engineering [21,22] and electronic discovery in legal proceedings [23]. In addition, several recent challenges, such as TREC [17,24] and CLEF eHealth task 2 [25–27], further promote the development of automatic document screening. To the best of our knowledge, two main research branches can be observed in the literature: information retrieval and machine learning.

In the area of information retrieval, the related investigations can be further divided into three groups: relevant feedback [20,28,29], query expansion [29–31] and ranking learning [32–34]. The former two methods emphasize transforming or improving the original query. The main difference is that relevant feedback is devoted to gathering information representing the user's need and automatically creating a new query, and query expansion reformulates a given query with synonyms or semantically related terms to match additional documents. The ranking learning methods sort all documents so that the relevant documents are ranked before irrelevant ones as many as possible.

In fact, document screening can also be regarded as a binary classification problem (*relevant* versus *nonrelevant*). In theory, any supervised machine learning model for text classification can be utilized directly, such as naïve Bayes [35], support vector machines [36,37], random forests [37] and so on. However, due to the severe imbalance of relevant and nonrelevant instances, time-consuming annotation and heavy workload, the performance of many supervised models is not satisfactory. In recent years, considerable effort has been spent on screening documents with *active learning* strategies [38–40]. The main idea of this strategy is that a supervised model can perform better with fewer annotated instances if it is allowed to choose the instances from which it learns [41]. It has been shown that this active learning solution outperforms its counterparts in many real-world cases [17,24–27]. Therefore, we adopted an active learning algorithm to screen scholarly articles about the origins of SARS-CoV-2.

2.2. Active learning

In many real-world applications, large numbers of unannotated instances are easily available, but annotated instances are time-consuming and expensive to obtain. In such a scenario, a machine learning algorithm can actively query an oracle (e.g., a human annotator) for the label of a focal instance. This type of iterative supervised learning method is called *active learning* [41]. It is sometimes referred to as *optimal experimental design* or *query learning* in the statistics literature [42]. The overall goal is to construct a classifier as good as possible with fewer labeled instances than necessary [43].

Active learning mainly consists of five steps, as illustrated in Fig 1. Given an unlabeled set S_1, these steps will be described briefly as follows.

Step 1 A labeled training set is initialized to an empty set, i.e., $S_2 \leftarrow \emptyset$.

Step 2 One *query strategy* is utilized to select the most valuable instance I_{ulab} from S_1, and then a label is assigned by an oracle to this instance I_{ulab}.

Step 3 The instance I_{ulab} with its label is added to the training set, viz. $S_2 \leftarrow S_2 \cup \{I_{ulab}\}$ and removed from the unlabeled set, viz. $S_1 \leftarrow S_1 - \{I_{ulab}\}$.

Step 4 A supervised machine learning model is retrained on the updated set S_2.

Step 5 Steps 2–4 are looped until a stopping criterion is met.

In Step 2, a variety of query strategies are put forward in the literature, such as uncertain sampling [44], expected error reduction [45], and query by committee [46,47]. The uncertain sampling query strategy, as its name implies, selects the instance on whose label the classifier is most uncertain. The expected error reduction strategy is devoted to annotating an instance so that the current classifier can achieve a lower generalization error. Different from the former two strategies, the query by committee strategy simultaneously considers multiple classifiers (viz. a committee of classifiers) and operates by querying the label of the instance on which

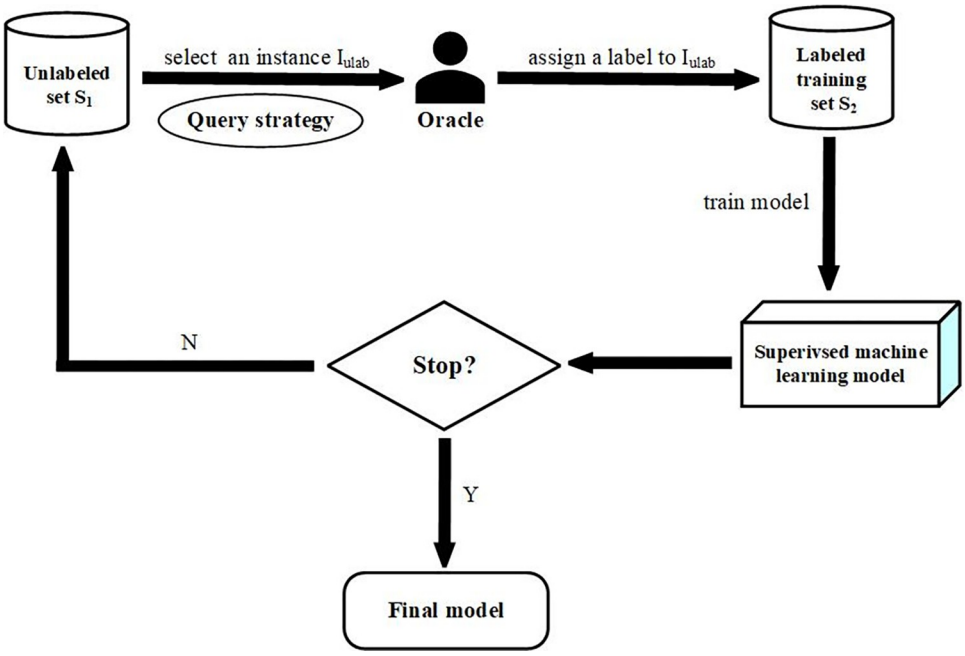

Fig 1. The procedure of an active learning approach.

https://doi.org/10.1371/journal.pone.0273725.g001

these classifiers disagree the most. Santos et al. [48] comprehensively compared the pros and cons of these query strategies on a large number of datasets and suggested that uncertain sampling and expected error reduction strategies should be preferred in many real-world scenarios.

In Step 3, a supervised machine learning model is involved. Theoretically speaking, any supervised classification model can be used in this step. However, due to different performance, time and space complexities, the following models were deployed in previous studies: support vector machine (SVM) [49–51], random forest (RF) [52], naïve Bayes classifier [48,53], and logistic regression (LR) [48,53].

In Step 4, the most important thing is when to stop this iterative procedure. Cormack et al. [23] argued that enough annotated instances should be seen as a signal to stop learning. In fact, it is usually very difficult to determine how many annotated instances are sufficient in real-world applications. Therefore, many scholars have considered whether a focal model approaches stability in terms of performance as a stopping criterion [43,54]. The measures for performance include the F_1 score, the area under the receiver operator characteristic (ROC) curve or the precision-recall (PR) curve.

3. Datasets

3.1. Data collection and annotation

Data collection. Due to the difficulty and complexity of the traceability of SARS-CoV-2, the available scientific publications are very scarce in the literature. Domingo [55] found only 1,675 results in the PubMed database with the search strategy "Origin of SARS-CoV-2" on July 19, 2021, but fewer than 100 articles disclosed scientific evidence about the origins of SARS--CoV-2. As of September 27, 2021, there are nearly eight million scholarly articles in the CORD-19 dataset [10]. In other words, it is very difficult and time-consuming to screen scholarly articles about the origins of SARS-CoV-2 from a large amount of literature.

However, to alleviate the workload of an oracle in active learning and to smoothly run active learning, this study aims to prepare a *seed* dataset of annotated publications in advance. Fortunately, the "WHO-convened Global Study of Origins of SARS-CoV-2: China Part" [56] and several review articles about the origins of SARS-CoV-2, such as [55–60], provide valuable clues. The general idea is to determine a small collection of seed articles in the first place and then expand it on the basis of forward and backward citations of these seed articles.

More specifically, once [55–60] are chosen, the following steps are conducted on each article in this dataset to determine our seed articles. (a) The forward and backward citations are retrieved from the Dimensions database [61] with the Dimensions API according to the resulting DOI (digital object identifier) [62]. (b) The metadata information of each citation (such as title, abstract, publication time, publication venue, and so on) is fetched from the PubMed database with EFetch API after mapping DOI to PMCID or PMID. (c) The noisy citations are removed with three manually curated rules: the publication year must be later than December 2019, the research topic should be related to COVID-19, and the resulting article must have been peer-reviewed.

To intuitively illustrate the rationale of our idea of collecting seed articles, we take a partial list of backward citations (references) in [57], shown in Fig 2, as an illustrative example. During the collection procedure, the following three rules are at work. First, since COVID-19 pneumonia broke out in December 2019, the articles published before December 2019 should not be related to the origins of SARS-CoV-2. Therefore, articles marked in yellow in Fig 2 are filtered out, such as (3), (6), (7), (8), (9), (15), and (17). Second, whether a scientific publication serves as a relevant instance or a nonrelevant instance, it should discuss COVID-19-related themes. Hence, one can rule out (13) in Fig 2. Last but not least, to focus on science, this study only considers peer-reviewed articles. In this way, preprints are excluded from further analysis, such as (1) in Fig 2. It is worth mentioning that we keep an eye on the status of each preprint by preprint-publication links [63]. Once it is published in a peer-reviewed venue, we will include it in our dataset. For the example in Fig 2, our seed dataset consists of (2), (4), (5), (10), (11), (12), (14), and (16).


(1) Dicken, S.J., Murray, M.J., Thorne, L.G., Reuschl, A.-K., Forrest, C., Ganesha-lingham, M., Muir, L., Kalemera, M.D., Palor, M., McCoy, L.E., et al. (2021). Characterisation of B.1.1.7 and Pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2021.03.22.436468.

(2) Dinnon, K.H., 3rd, Leist, S.R., Schäfer, A., Edwards, C.E., Martinez, D.R., Montgomery, S.A., West, A., Yount, B.L., Jr., Hou, Y.J., Adams, L.E., et al. (2020). A mouse-adapted model of SARS-CoV-2 to test COVID-19 countermeasures. Nature *586*, 560–566.

(3) Follis, K.E., York, J., and Nunberg, J.H. (2006). Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. Virology *350*, 358–369.

(4) Freuling, C.M., Breithaupt, A., Müller, T., Sehl, J., Balkema-Buschmann, A., Rissmann, M., Klein, A., Wylezich, C., Höper, D., Wernike, K., et al. (2020). Susceptibility of raccoon dogs for experimental SARS-CoV-2 infection. Emerg. Infect. Dis. *26*, 2982–2985.

(5) Gallaher, W.R. (2020). A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Arch. Virol. *165*, 2341–2348.

(6) Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y., and Shi, Z. (2012). Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. J. Virol. *86*, 4620–4630.

(7) Ge, X.-Y., Li, J.-L., Yang, X.-L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., Peng, C., et al. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature *503*, 535–538.

(8) Geddes, A.M. (2006). The history of smallpox. Clin. Dermatol. *24*, 152–157.

(9) Gombold, J.L., Hingley, S.T., and Weiss, S.R. (1993). Fusion-defective mutants of mouse hepatitis virus A59 contain a mutation in the spike protein cleavage signal. J. Virol. *67*, 4504–4512.

(10) Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al.; Sheffield COVID-19 Genomics Group (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell *182*, 812–827.e19.

(11) Kuzmina, A., Khalaila, Y., Voloshin, O., Keren-Naus, A., Boehm-Cohen, L., Raviv, Y., Shemer-Avni, Y., Rosenberg, E., and Taube, R. (2021). SARS-CoV-2 spike variants exhibit differential infectivity and neutralization resistance to convalescent or post-vaccination sera. Cell Host Microbe *29*, 522–528.e2.

(12) Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., et al. (2020). Origin and cross-species transmission of bat coronaviruses in China. Nat. Commun. *11*, 4235.

(13) Lednicky, J.A., Tagliamonte, M.S., White, S.K., Elbadry, M.A., Alam, M.M., Stephenson, C.J., Bonny, T.S., Loeb, J.C., Telisma, T., Chavannes, S., et al. (2021). Emergence of porcine delta-coronavirus pathogenic infections among children in Haiti through independent zoonoses and convergent evolution. medRxiv. https://doi.org/10.1101/2021.03.19.21253391.

(14) Leist, S.R., Dinnon, K.H., 3rd, Schäfer, A., Tse, L.V., Okuda, K., Hou, Y.J., West, A., Edwards, C.E., Sanders, W., Fritch, E.J., et al. (2020). A mouse-adapted SARS-CoV-2 induces acute lung injury and mortality in standard laboratory mice. Cell *183*, 1070–1085.e12.

(15) Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

(16) Li, L.-L., Wang, J.-L., Ma, X.-H., Sun, X.-M., Li, J.-S., Yang, X.-F., Shi, W.-F., and Duan, Z.-J. (2021). A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province, China. Emerg. Micro. Infect. *10*, 1683–1690.

(17) Lim, P.L., Kurup, A., Gopalakrishna, G., Chan, K.P., Wong, C.W., Ng, L.C., Se-Thoe, S.Y., Oon, L., Bai, X., Stanton, L.W., et al. (2004). Laboratory-acquired severe acute respiratory syndrome. N. Engl. J. Med. *350*, 1740–1745.

Fig 2. An illustrative example with a partial list of backward citations in [57].

https://doi.org/10.1371/journal.pone.0273725.g002

Table 1. The number of articles in the top 10 journals.

Journal	# of articles	Journal	# of articles
Journal of Medical Virology	18	Nature Communications	9
Nature	14	Lancet	7
Science	14	Infection, Genetics and Evolution	6
Cell	10	National Science Review	6
Emerging Microbes & Infections	9	Scientific Reports	6

https://doi.org/10.1371/journal.pone.0273725.t001

Ultimately, this work collected 470 articles in total from 282 journals, covering the PubMed, Elsevier, and WHO databases. The involved fields include title, abstract, journal/conference, publication time and DOI. Table 1 shows the top 10 journals in our seed dataset, where Journal of Medical Virology ranks first in terms of the number of articles, followed by Nature, Science and Cell.

Data annotation. Once our seed articles are determined, we need to attach a *relevant* or *nonrelevant* label to each article for active learning. Two annotators majoring in biology independently annotated all publications by reading the abstract and main body of every article. These two annotators were from the College of Biological Sciences and Biotechnology, Beijing Forestry University. Their research interests include the transmission and prevention of coronavirus. Furthermore, they have recently annotated the entities mentioned in the articles on COVID-19. Hence, they should be competent for the annotation work of our experiment.

To accurately annotate the articles in the seed dataset, we design an annotation guideline. It mainly gives several suggestions on which articles should be labeled as *relevant* or *nonrelevant*. The whole annotation process is mainly divided into two stages. In the first stage, to unify their understanding of the guideline, 50 of the same articles are assigned to these two annotators. The interannotator agreement is calculated with the multi-κ indicator [64]. The agreement between the two annotators was 80.2%. On closer examination, we find that the annotators have a different understanding of the articles mentioning intermediate hosts of SARS-CoV-2 (such as ferrets, cats, and dogs). Through extensive discussions, we argue that such articles should be relevant to the origins of SARS-CoV-2. Thereupon, the guideline is correspondingly revised. Then, according to the updated guidelines, they independently annotated the remaining articles in the second round as the final annotation results.

Ultimately, our annotated corpus comprises 170 relevant articles (positive instances) and 300 nonrelevant articles (negative instances). Their trends with publishing time are shown in Fig 3. The publication time ranged from December 2019 to October 2021. Most articles were published between May 2020 and October 2020. The number of relevant and nonrelevant articles peaked in May 2020 and October 2020, respectively. These trends in Fig 3 are basically in line with the global trends in COVID-19 research [65].

3.2. Document representation

Another critical ingredient for screening scholarly articles about the origins of SARS-CoV-2 is how to represent a document with a fixed-length vector for active learning. Although many document representation methods are put forward in the literature, such as extensions of words to documents [66], convolution-based methods [67], and variational autoencoders [68], they are not able to leverage citation information between scientific documents. This greatly limits their representation power at the document level. Cohan et al. [69] developed a novel document representation approach, namely, SPECTER (Scientific Paper Embedding using Citation-informed TransforERs), through pretraining a transfer language model on the

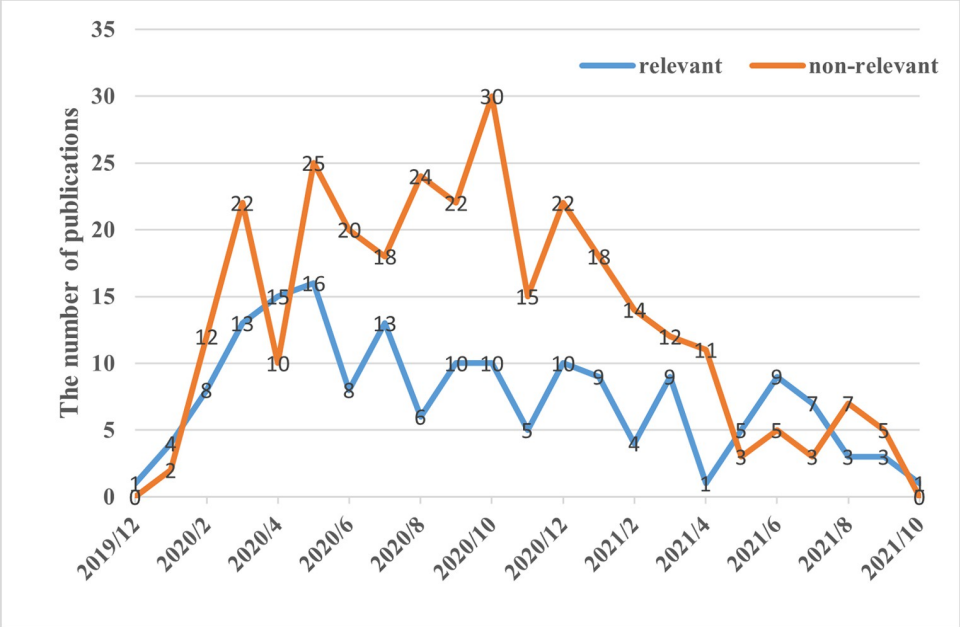

Fig 3. The trends of the number of relevant and nonrelevant publications with publishing time in our seed dataset.

https://doi.org/10.1371/journal.pone.0273725.g003

citation network of scientific documents. Thus, no task-specific fine-tuning is needed for our task, so this work prefers the SPECTER method.

It is worth noting that document embeddings with the SPECTER method on the basis of titles, abstracts and citation network are also released with each CORD-19 update [10]. More specifically, each scientific publication is represented with a 768-dimensional dense vector. To obtain these representations, we map each document in the seed dataset to that in the CORD-19 dataset through the resulting DOI [62].

4. Methods

To screen scholarly articles about the origins of SARS-CoV-2, our research framework, as shown in Fig 4, mainly consists of four modules. After collecting and labeling seed articles (cf. Subsection 3.1), we retrieve document representations of these articles from the CORD-19 dataset (cf. Subsection 3.2). Then, an active learning-based approach with SVM, LR or RF as a base classifier is deployed after optimizing the query strategy. On the basis of three tuned models with an active learning strategy, scholarly articles about the origins of SARS-CoV-2 are

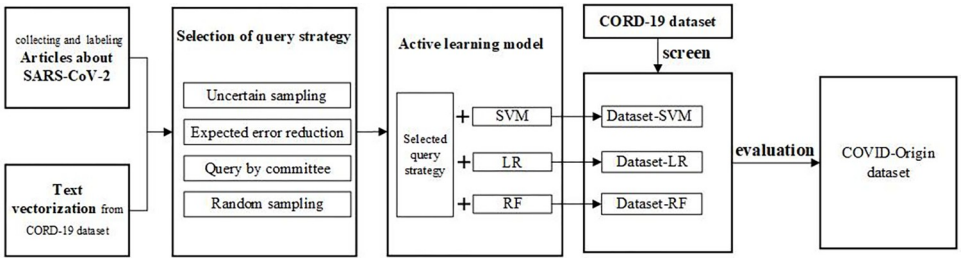

Fig 4. Research framework for screening articles about the origins of SARS-CoV-2.

https://doi.org/10.1371/journal.pone.0273725.g004

screened from the CORD-19 dataset and checked manually one by one for the top 1000 documents from each base classifier. In the end, a dataset about the origins of SARS-CoV-2, named the COVID-Origin dataset, is constructed. In the following subsections, the last three modules are described at length.

4.1. Query strategy

1) Uncertain sampling. This query strategy selects the most uncertain instance for labeling. The most uncertain instance is referred to as the instance that the current classifier is most likely to make a mistake. Intuitively, such an instance can improve the performance of the model more efficiently. The uncertainty of an instance can be measured by information entropy. The more uncertainty an instance has, the greater its information entropy [70] is. Formally, this strategy can be expressed as follows:

$$x^* = \text{argmax}_x - [\Pr(\text{relevant}|x) \log \Pr(\text{relevant}|x) + \Pr(\text{non\_relevant}|x) \log \Pr(\text{non\_relevant}|x)]$$

Here, Pr(relevant|x) and Pr(non_relevant|x) represent the probability of x being classified into relevant and nonrelevant categories, respectively. When these probabilities approach 0.5, the resulting instance will be more likely to be selected.

2) Expected error reduction. This strategy first estimates the generalization error of the current classifier and then sequentially evaluates the generalization error change that may be brought to the classifier if a new instance is added to the training set. Finally, it selects the instance for labeling that can reduce the expected generalization error the most. It is the generalization error minimization that enables this strategy to become an effective query strategy [71]. Nevertheless, this strategy brings a huge time cost due to its error reduction estimation, so it is inefficient for active learning on a large-scale dataset. Therefore, several approximate alternatives are proposed in the literature [71,72]. To speed up the process, the approximated error reduction in [72] is utilized in this study.

3) Query by committee. In this strategy, multiple classifiers, namely, a committee of classifiers, are involved. The instance on which these classifiers disagree the most by voting will be chosen for labeling. The evaluation criteria for committee voting include entropy, Kullback–Leibler divergence, and Jensen–Shannon divergence. For simplicity, this study adopts voting entropy, which is defined formally as follows:

$$x^* = \text{argmax}_x - \left[\frac{V(x, \text{relevant})}{M} \log \frac{V(x, \text{relevant})}{M} + \frac{V(x, \text{non\_relevant})}{M} \log \frac{V(x, \text{non\_relevant})}{M}\right]$$

Here, $V(x, \text{relevant})$ and $V(x, \text{non\_relevant})$ are the number of votes of the committee classifying instance x into relevant and nonrelevant categories, respectively. M is the total number of classifiers in a committee. When the votes of relevant and nonrelevant categories are approximately evenly distributed, the resulting instance will be more likely to be chosen.

4.2. Candidates in the CORD-19 dataset

Once our active learning-based approach with a proper query strategy is developed, it will be utilized to screen scientific publications from the CORD-19 dataset for follow-up real-world applications. In fact, the CORD-19 dataset covers scholarly articles on MERS-CoV, SAR-CoV and SARS-CoV-2. Therefore, before screening scholarly articles on the origins of SARS-CoV-2, articles that are not related to COVID-19 should be eliminated in advance.

In more detail, the following two steps are conducted. (1) This work extracts articles containing "COVID-19", "2019-nCoV", "SARS-CoV-2" or "coronavirus 2019" in the title or abstract. (2) The seed publications in Subsection 3.1 are excluded from the subset from the

previous step. After these two steps, the candidates on the origins of SARS-CoV-2 are obtained for further screening with our active learning-based approach. For convenience, this subset is denoted as the *CORD-19 subset*.

4.3. Screening procedure

For ease of understanding, the pseudocode of our methodology is summarized in Algorithm 1. Our input includes the initial labeled set S, unlabeled set S_1, query strategy, base classifier, and CORD-19 subset. Our algorithm mainly consists of the following three parts. (1) The classifier f is initialized with the labeled training set S_2 in the first place (Line 3–4). (2) After an instance I_{ulab} chosen by the query strategy is annotated by an oracle, deleted from unlabeled set S_1 and added to the labeled training set S_2, the classifier f is retrained on the updated S_2. This procedure is repeated until f reaches the best performance (Lines 5–10). (3) Finally, the tuned classifier f is utilized to screen the articles on the origins of SARS-CoV-2 from the CORD-19 subset (Lines 11–12).

Algorithm 1. Algorithms for screening articles about the origins of SARS-CoV-2.

1: **Input:** initial labeled set S, unlabeled set S_1, query strategy, classifier f, and CORD-19 subset
2: **Process:**
3: Initialize the labeled training set $S_2 = S$
4: Train an initial classifier f with the labeled training set S_2
5: **Repeat**
6: Generate one new instance I_{ulab} from the unlabeled set S_1 according to the query strategy
7: Get the label of I_{ulab} from an oracle
8: Update the labeled training set S_2 and unlabeled set S_1
9: Update the classifier f with labeled training set S_2
10: **Until** the best performance of the classifier f is reached
11: Screen the publications on the origins of SARS-CoV-2 from the CORD-19 subset
12: **Output:** articles about the origins of SARS-CoV-2

5. Experimental results and discussions

The Python toolkit ALiPy [73] implements more than 20 commonly used active learning methods. Hence, it is utilized to screen articles on the origins of SARS-CoV-2. It is noteworthy that since a seed dataset of annotated publications is prepared ahead (cf. Subsection 3.1), our experiments are performed by simulating the labeling process by an oracle. That is, the resulting label of the document chosen by a query strategy is assumed to be unknown beforehand and must be assigned by an oracle during the active learning procedure.

5.1. Query strategy optimization

To tune the query strategy, our seed dataset is split randomly into a training set and a test set with a ratio of 7:3 and a similar relevant/nonrelevant distribution. That is, our training and test sets are made up of 329 and 141 instances, respectively. As shown in Algorithm 1, the base classifier f needs to be initialized. For this purpose, 8 instances are selected randomly from the training set as the initial labeled set S. Thus, our unlabeled set S_1 is composed of the remaining 321 instances in the training set. In addition, apart from three query strategies (cf. Subsection 4.1), a random sampling query strategy is also used in this study. In fact, this query strategy is equivalent to the traditional supervised classification approach.

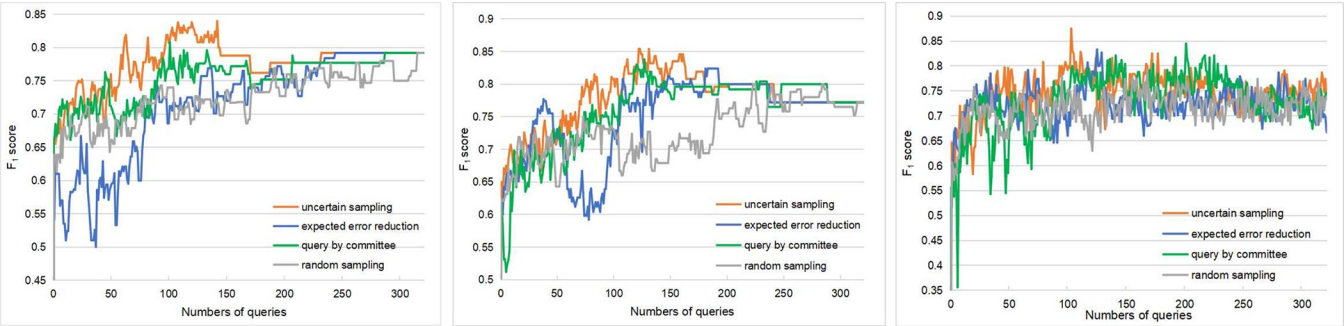

Fig 5. The performance of our active learning approach with support vector machine (a), logistic regression (b), and random forest (c) as the respective base classifiers in terms of the F_1 score.

https://doi.org/10.1371/journal.pone.0273725.g005

This study considers three base classifiers: support vector machines (SVM), logistic regression (LR), and random forest (RF). Fig 5 illustrates the performance of our active learning approach with different base classifiers and different query strategies on the test set in terms of the F_1 score. From Fig 5, it is obvious that the active learning approach converges faster than the traditional supervised classification counterpart (viz. active learning approach armed with a random sampling query strategy). Among the three commonly used query strategies, the active learning approach armed with an uncertain sampling query strategy has the best performance, followed by the active learning approach armed with a query by committee query strategy. For base classifiers, the active learning approach with random forest (RF) as a base classifier has stable performance, regardless of the query strategy used.

To select an appropriate query strategy, this work simultaneously considers the F_1 score and the number of queries when the active learning approach performs best on the test data, as reported in Table 2. Note that the number of queries is utilized to measure the workload saved by the active learning approach. A lower value indicates more workload saved. From Table 2, it is apparent that the performance of the active learning approach armed with an uncertain sampling query strategy is better than that of the active learning approach armed with other query strategies in terms of the F_1 score and the number of queries. For example, after 103 queries, the combination of an uncertain sampling query strategy with random forest (RF)

Table 2. F1 score and numbers of queries when each model reached the maximum F1 score.

Base classifier	Query strategy	F_1 score	# of annotated articles	
			Relevant	Nonrelevant
SVM	uncertain sampling	**0.840**	65	77
	expected error reduction	0.792	112	132
	query by committee	0.808	**39**	**62**
	random sampling	0.792	110	205
LR	uncertain sampling	**0.854**	**43**	**79**
	expected error reduction	0.824	67	113
	query by committee	0.838	49	77
	random sampling	0.799	81	142
RF	uncertain sampling	**0.875**	**47**	**56**
	expected error reduction	0.833	59	66
	query by committee	0.845	90	111
	random sampling	0.791	65	102

https://doi.org/10.1371/journal.pone.0273725.t002

achieves the best F_1 score. That is, only 103 labeled articles (except for 8 articles for initializing a base classifier) are needed to reach the maximum F_1 score instead of all annotated articles. This indicates that approximately two-thirds of the workload for annotating articles can be saved. Hence, the active learning approach armed with an uncertain sampling query strategy is further used for screening scholarly articles about the origins of SARS-CoV-2 from the CORD-19 dataset in the next subsection.

5.2. Screening articles about the origins of SARS-CoV-2

To screen articles about the origins of SARS-CoV-2, a comprehensive literature dataset, CORD-19 (2021-9-27 version), is utilized here. According to the criteria in Subsection 4.2, we can obtain a CORD-19 subset, which consists of 371,664 candidates in total. Then, our active learning approach with SVM, LR or RF as a base classifier independently assigns a posterior probability of the relevant category to each candidate. On the basis of posterior probabilities, the top 1,000 articles of each base classifier were checked manually one by one. This procedure is very similar to the annotation process in Subsection 3.1. Note that the articles in the CORD-19 dataset [10] come from multiple sources, such as WHO's COVID-19 database, PubMed Central, MedLine, Elsevier and so on. To deduplicate publications, a conservative clustering policy in which any identifier (such as *doi*, *pmc_id*, *pubmed_id*, *arxiv_id*, *who_covidence_id*, and *mag_id*) conflict prohibits clustering was utilized. This enables many duplicative articles to appear in the CORD-19 dataset. This study further clusters these articles if any identifier matches and manually checks top articles in terms of posterior probabilities. Here, the top 1,000 articles actually correspond to the top ~1,800 articles in the original dataset.

In this way, we can evaluate the performance of our approach in terms of accuracy, as depicted in Fig 6. The left vertical axis denotes the accuracy, and the right vertical axis is the number of relevant articles. Among the top 1,000 articles, the SVM, LR and RF base classifiers correctly screened 425, 465, and 489 articles, respectively. As the posterior probability of the relevant category decreases, the accuracy of the screened articles shows a downward trend. This is in line with our intuition. We take the top 200 articles as an example. The accuracies of

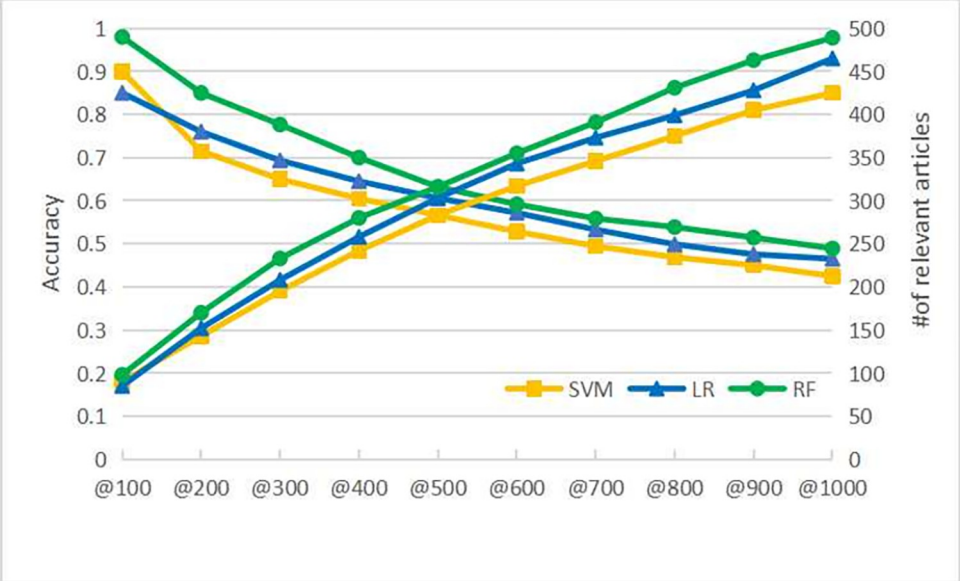

Fig 6. The performance of the top screened articles about the origins of SARS-CoV-2 from the CORD-19 dataset.

https://doi.org/10.1371/journal.pone.0273725.g006

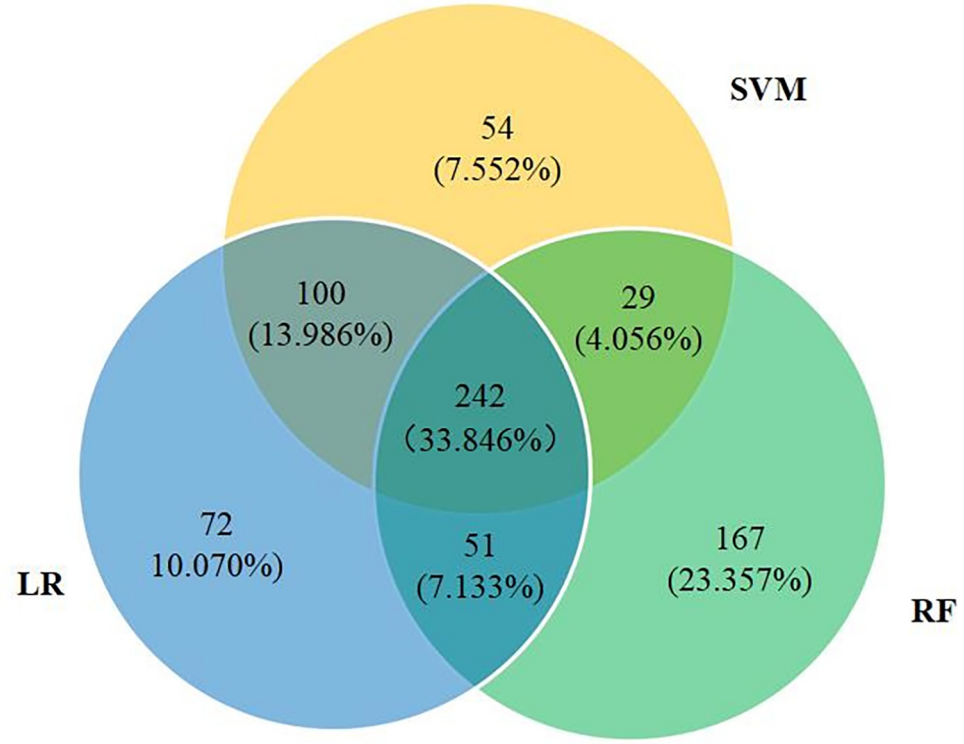

Fig 7. The overlapping shares of relevant articles screened by three base classifiers.

https://doi.org/10.1371/journal.pone.0273725.g007

all three base classifiers reach more than 70%, and the accuracy of the RF classifier even exceeds 80%. This indicates that our active learning-based approach for screening articles about the origins of SARS-CoV-2 is feasible.

In total, 715 unique articles were screened among the top 1,000 scholarly articles by three base classifiers. That is, the articles screened by three base classifiers overlap greatly. Fig 7 depicts the overlapping shares of relevant articles screened by three base classifiers. It is not difficult to see that the articles screened simultaneously by three base classifiers account for 33.846%, and those screened by two classifiers account for at least 59.021%. This indicates that each base classifier has its pros and cons and cannot serve as an alternative to the others. In real-world applications, it is better to screen scientific publications on the origins of SARS--CoV-2 with multiple base classifiers in our framework (cf. Fig 4).

6. Conclusions

The outbreak of COVID-19 has disrupted people's daily lives and work for nearly two years. To completely solve this epidemic, one of the most important tasks is to trace the origins of SARS-CoV-2. Due to the complexity of traceability work, the origins of SARS-CoV-2 are still inconclusive. It is necessary for researchers to build a full picture of previous studies on the origins of SARS-CoV-2 in advance and then to conduct further investigations. However, there is currently no comprehensive literature dataset on the origins of SARS-CoV-2 that can be used by researchers. Therefore, to bridge this gap, this study is devoted to screening scholarly articles about the origins of SARS-CoV-2 from large numbers of scientific publications.

For this purpose, we propose an active learning-based approach that can quickly screen articles with better accuracy and save the labeling workload of human annotators. In more

detail, after collecting and labeling a small seed dataset of articles, we develop the active learning-based approach with three query strategies and three base classifiers (SVM, RF, and LR). Extensive experiments show that our approach has better performance than its traditional counterparts, and the uncertain sampling query strategy performs best among the three strategies. To quantitatively evaluate the performance of the three base classifiers, we manually checked the top 1,000 articles one by one in terms of posterior probabilities. In the end, three classifiers screened 425, 465 and 489 relevant articles. In total, there were 715 unique articles, more than 50% of which were screened by at least two base classifiers.

However, there is still room to improve our approach. For example, only three query strategies are taken into consideration in this work. In the near future, other query strategies, such as the graph density query strategy [74], will be used to screen scholarly articles on the origins of SARS-CoV-2. In addition, due to the pros and cons of each base classifier, ensemble learning will be utilized in our next work as a base classifier for our active learning-based approach.

Author Contributions

Conceptualization: Xin An, Shuo Xu.

Data curation: Mengmeng Zhang.

Methodology: Xin An, Shuo Xu.

Validation: Xin An.

Visualization: Mengmeng Zhang.

Writing – original draft: Mengmeng Zhang.

Writing – review & editing: Xin An, Shuo Xu.

References

1. Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2010; 579(7798): 265–269.
2. Relman DA. Opinion: To Stop the next pandemic, we need to unravel the origins of COVID-19. Proceedings of the National Academy of Sciences of the United States of America. 2020; 117(47): 29246–29248. https://doi.org/10.1073/pnas.2021133117 PMID: 33144498
3. Wang N, Li SY, Yang XL, Huang HM, Zhang YJ, Guo H, et al. Serological Evidence of Bat SARS-related Coronavirus Infection in Humans, China. Virologica Sinica. 2018; 33: 104–107. https://doi.org/10.1007/s12250-018-0012-7 PMID: 29500691
4. Weingartl HM., Nfon C, and Kobinger G. Review of Ebola virus infections in domestic animals. Developments in Biologicals. 2013; 135: 211–218. https://doi.org/10.1159/000178495 PMID: 23689899
5. da Silva J. A. T, Tsigaris P, Erfanmanesh M. Publishing volumes in major databases related to Covid-19. Scientometrics. 2021; 126(1): 831–842. https://doi.org/10.1007/s11192-020-03675-3 PMID: 32904414
6. Chen Q, Allot A, Lu Z. Keep up with the coronavirus research. Nature. 2020; 579(7798): 193–193.
7. Wang X, Song X, Li B, Guan Y, Han J. Comprehensive Named Entity Recognition on CORD-19 with distant or weak supervision. ArXiv: abs/2003.12218. 2020. Available from: https://arxiv.org/abs/2003.12218.
8. Zuo X, Chen Y, Ohno-Machado L, Xu H. How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. Briefings in Bioinformatics. 2021; 22(2): 800–811. https://doi.org/10.1093/bib/bbaa331 PMID: 33757278
9. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research. 2021; 49(D1): D1534–D1540. https://doi.org/10.1093/nar/gkaa952 PMID: 33166392
10. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: The COVID-19 Open Research Dataset. Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020; arXiv:2004.10706v2. PMID: 32510522.

11. Xu S, Hao L, An X, Pang H, Li T. Review on emerging research topics with key-route main path analysis. Scientometrics. 2020; 122: 607–624.

12. Sinatra R, Deville P, Szell M, Wang D, Barabási AL. A century of physics. Nature Physics. 2015; 11: 791–796.

13. Blasius B. Power-law distribution in the number of confirmed covid-19 cases. Chaos. 2020; 30(9): 093123. https://doi.org/10.1063/5.0013031 PMID: 33003939

14. Giacomet V, Stracuzzi M, Paradiso L, Di Cosiom ME, Rubinacci V, Zuccotti G. Defining the clinical phenotype of COVID-19 in children. Pediatric Allergy and Immunology. 2020; 31(s26):82–84. https://doi.org/10.1111/pai.13355 PMID: 33236440

15. Lam T, Jia N, Zhang Y Shum M H, Jiang J, Zhu H, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan Pangolins. Nature. 2020; 583(7815):282–285. https://doi.org/10.1038/s41586-020-2169-0 PMID: 32218527

16. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research. 2020; 176:104742. https://doi.org/10.1016/j.antiviral.2020.104742 PMID: 32057769

17. Grossman MR, Cormack GV, Roegiest A. TREC 2016 Total Recall Track Overview. Proceedings of the 25th Text REtrieval Conference (TREC 2016). 2016; http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf.

18. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Annals of Internal Medicine. 1997; 127(5): 380–387. https://doi.org/10.7326/0003-4819-127-5-199709010-00008 PMID: 9273830

19. Carvallo A, Parra D, Lobel H, Soto A. Automatic document screening of medical literature using word and text embeddings in an active learning setting. Scientometrics. 2020; 125(3):3047–3084.

20. Garc Adeva JJ, Pikatza Atxa JM, Ubeda Carrillo M, Ansuategi Zengotitabengoa E. Automatic text classification to support systematic reviews in medicine. expert systems with applications. 2014; 41(4): 1498–1508.

21. Hassler EE, Hale DP, Hale JE. A comparison of automated training-by-example selection algorithms for evidence based software engineering. Information and Software Technology. 2018; 98: 59–73.

22. Yu Z, Kraft NA, Menzies T. Finding better active learners for faster literature reviews. Empirical Software Engineering. 2018; 23(6): 3161–3186.

23. Cormack GV, Grossman MR. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2014; 153–162.

24. Roegiest A, Cormack GV, Grossman MR, Clarke CLA. TREC 2015 Total Recall Track Overview. Proceedings of the 24th Text REtrieval Conference (TREC 2015). 2015; https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf.

25. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2017 technologically assisted reviews in empirical medicine overview. CEUR Workshop Proceedings. 2017; 1866.

26. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2018 technologically assisted reviews in empirical medicine overview. CEUR Workshop Proceedings. 2018; 2125.

27. Evangelos K, Dan L, Leif A, Ren S. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. CEUR Workshop Proceedings. 2019; http://ceur-ws.org/Vol-2380/paper_250.pdf.

28. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. International journal of computational biology and drug design. 2013; 6(1–2):5–17. https://doi.org/10.1504/IJCBDD.2013.052198 PMID: 23428470

29. Donoso-Guzmán I, Parra D. An interactive relevance feedback interface for evidence-based health care. In 23rd international conference on intelligent user interfaces. 2014;103–114. https://doi.org/10.1145/3172944.3172953.

30. Yang Y, Bansal N, Dakka W, Ipeirotis P, Koudas N, Papadias D. Query by document. Proceedings of the Second ACM International Conference on Web Search and Data Mining. 2009; 34–43. https://doi.org/10.1145/1498759.1498806.

31. Weng L, Li ZW, Cai R, Zhang YX, Zhou YZ, Yang LT, et al. Query by document via a decomposition-based two-level retrieval approach. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011;505–514. https://doi.org/10.1145/2009916.2009985.

32. Lee GE, Sun A. Seed-driven document ranking for systematic reviews in evidence-based medicine. In The 41st international ACM SIGIR conference on research & development in information retrieval. 2018; 455–464. https://doi.org/10.1145/3209978.3209994.

33. Goodwin TR, Harabagiu SM. Knowledge representations and inference techniques for medical question answering. In ACM transactions on intelligent systems and technology (TIST). 2018; 9(2) 2157–6904.
34. Grotov A, de Rijke M. Online learning to rank for information retrieval. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016; 4: 1215–1218.
35. Xu S. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science. 2018; 44 (1): 48–59.
36. Xu S, An X, Qiao X, Zhu L. Multi-task least-squares support vector machines. Multimedia Tools and Applications. 2014; 71(2): 699–715.
37. An X, Sun X, Xu S, Hao L, Li J. Important citations identification by exploiting generative model into discriminative model. Journal of Information Science. 2021. https://doi.org/10.1177/0165551521991034
38. Shi Y, Yao K, Tian L, Jiang D. Deep LSTM based feature mapping for query classification. Conference of north American chapter of the association for computational linguistics: Human language technologies. 2016:1501–1511.
39. Peters ME, Neumann M, Lyyer M, Gardner M. Deep contextualized word repressntations. North American of the associate on for computational linguistics. 2018: 2227–2237.
40. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. Environment International. 2020; 138:105623. https://doi.org/10.1016/j.envint.2020.105623 PMID: 32203803
41. Settles B. Active Learning Literature Survey. University of Wisconsin—Madison. 2009; http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.
42. Lewi J, Butera R, Paninski L. Sequential Optimal Design of Neurophysiology Experiments. Neural Computation. 2009; 21 (3): 619–687. https://doi.org/10.1162/neco.2008.08-07-594 PMID: 18928364
43. Settles B. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2012; 6 (1):1–114.
44. Tang M, Luo X, Rouko S. Active learning for statistical natural language parsing. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). 2002; 120–127.
45. Roy N, McCallum A. Toward optimal active learning through sampling estimation of error reduction. In Proceedings of the International Conference on Machine Learning (ICML). 2001; 441–448.
46. Seung HS, Oppert M, Sompolinsky, H. Query by Committee. Proceedings of the fifth annual workshop on computational learning theory. 1992; 287–294.
47. Melville P, Yang SM, Saar-Tsechansky M, Mooney R. Active Learning for Probability Estimation Using Jensen-Shannon Divergence. Berlin, Heidelberg, Springer Berlin Heidelberg. 2015; 268–279. https://doi.org/10.1007/11564096_28.
48. dos Santos DP, Prudêncio R, Carvalho A. Empirical investigation of active learning strategies. Neurocomputing. 2019; 326–327: 15–27.
49. Yu Z, Menzies T. Total recall, language processing, and software engineering. Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering. 2018;10–13. https://doi.org/10.1145/3283812.3283818.
50. Huang S, Jin R, Zhou Z. Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014; 36(10):1936–1949. https://doi.org/10.1109/TPAMI.2014.2307881 PMID: 26352626
51. Gavves E, Mensink T, Tommasi T, Snoek CGM, Tuytelaars T. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. 2015 IEEE International Conference on Computer Vision (ICCV). 2015; 2731–2739. https://doi.org/10.1109/ICCV.2015.313.
52. Josu M, Borja A, Manuel G. Random forest active learning for AAA thrombus segmentation in computed tomography angiography images. Neurocomputing. 2014; 126:71–77.
53. Ramirez-Loaiza ME, Sharma M, Kumar G, Bilgic M. Active learning: an empirical study of common baselines. Data Mining and Knowledge Discovery. 2017; 31:287–313. https://doi.org/10.1007/s10618-016-0469-7.
54. Byron CW, Thomas AT, Joseph L, Carla B, Christopher HS. Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics. 2010; 11:55. https://doi.org/10.1186/1471-2105-11-55 PMID: 20102628
55. Domingo JL. What we know and what we need to know about the origin of SARS-CoV-2. Environmental research. 2021; 200:111785. https://doi.org/10.1016/j.envres.2021.111785 PMID: 34329631

56. Joint WHO-China Study Team. WHO-convened global study of origins of SARS-CoV-2: China part. World Health Organisation: Geneva, Switzerland. 2021. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

57. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO et al. The origins of SARS-CoV-2: A critical review. Cell. 2021; 184(19):4848–4856. https://doi.org/10.1016/j.cell.2021.08.017 PMID: 34480864

58. van Helden J, Butler CD, Achaz G, Canard B, Casane D, Claverie JM et al. An appeal for an objective, open, and transparent scientific debate about the origin of SARS-CoV-2. Lancet. 2021; 398 (10309):1402–1404. https://doi.org/10.1016/S0140-6736(21)02019-5 PMID: 34543608

59. Karlsson EA, Duong V. The continuing search for the origins of SARS-CoV-2. Cell. 2021; 184(17): 4373–4374. https://doi.org/10.1016/j.cell.2021.07.035 PMID: 34416143

60. Leitner T, Kumar S. Where Did SARS-CoV-2 Come From? Molecular biology and evolution. 2020; 37 (9):2463–2464. https://doi.org/10.1093/molbev/msaa162 PMID: 32893295

61. Thelwall M. Dimensions: A Competitor to Scopus and the Web of Science? Journal of Informetrics, 2018, 12(2): 430–435.

62. Xu S, Hao L, An X, Zhai D, Pang H. Types of DOI errors of cited references in Web of Science with a cleaning method. Scientometrics. 2019; 120: 1427–1437. https://doi.org/10.1007/s11192-019-03162-4.

63. Cabanac G, Oikonomidi T, Boutron I. Day-to-day discovery of preprint–publication links. Scientometrics. 2021; 126:5285–5304. https://doi.org/10.1007/s11192-021-03900-7 PMID: 33897069

64. Davies M, Fleiss JL. Measuring Agreement for Multinomial Data. Biometrics. 1982; 38(4): 1047–1051.

65. Wang P, Tian D. Bibliometric analysis of global scientific research on COVID-19. Journal of biosafety and biosecurity. 2021; 3(1):4–9. https://doi.org/10.1016/j.jobb.2020.12.002 PMID: 33521590

66. Van Gysel C, De Rijke M, Kanoulas E. Neural Vector Spaces for Unsupervised Information Retrieval. ACM Transactions on Information Systems. 2018; 36(4):1–25.

67. Zamani H, Dehghani M, Croft WB, Learned-Miller E, Kamps J. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018; 497–506.

68. Wang W, Tao C, Gan Z, Wang G, Chen L, Zhang X. Improving Textual Network Learning with Variational Homophilic Embeddings. Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019;2076–2087.

69. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020; 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207.

70. Tang M, Luo X, Roukos S. Active Learning for Statistical Natural Language Parsing. Association for Computational Linguistics. 2002; 8:120–127.

71. Aodha OM, Campbell NDF, Kautz J, Brostow GJ. Hierarchical Subquery Evaluation for Active Learning on a Graph. Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition. 2014; 564–571. https://doi.org/10.1109/CVPR.2014.79.

72. Fu W, Wang M, Hao S, Wu X. Scalable Active Learning by Approximated Error Reduction. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018; 1396–1405. https://doi.org/10.1145/3219819.3219954.

73. Tang YP, Li GX, Huang SJ. ALiPy: Active Learning in Python. 2019; ArXiv: 1901.03802.

74. Ebert S, Fritz M, Schiele B. RALF: A reinforced active learning formulation for object class recognition. 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012; 3626–3633. https://10.1109/CVPR.2012.6248108.