Journal of Informetrics 17 (2023) 101376

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Deep representation learning of scientific paper reveals its potential scholarly impact

Zhuoren Jiang, Tianqianjin Lin, Cui Huang*

Department of Information Resources Management, School of Public Affairs, Zhejiang University, Hangzhou, 310058 PR China

ARTICLE INFO

Keywords:
Scholarly impact
Deep representation learning
Topicality
Originality

ABSTRACT

Citation and citation-based metrics are traditionally used to quantify the scholarly impact of scientific papers. However, for documents without citation data, i.e., newly published papers, the citation-based metrics are not available. By leveraging deep representation techniques, we propose a text-content based approach that may reveal the scholarly impact of papers without human domain-specific knowledge. Specifically, a large-scale Pre-Trained Model (PTM) with 110 million parameters is utilized to automatically encode the paper into the vector representation. Two indicators, $\boldsymbol{\tau}$ ***(Topicality)*** and $\boldsymbol{\sigma}$ ***(Originality)***, are then proposed based on the learned representations. These two indicators leverage the spatial relations of paper representations in the semantic space to capture the impact-related characteristics of a scientific paper. Extensive experiments have been conducted on a COVID-19 open research dataset with 1,056,660 papers. The experimental results demonstrate that the deep representation learning method can better capture the scientific content in the published literature; and the proposed indicators are positively and significantly associated with a paper's potential scholarly impact. In the multivariate regression analysis for the potential impact of a paper, the coefficients of σ and τ are 5.4915 ($P < 0.001$) and 6.6879 ($P < 0.001$) for next 6 months prediction, 12.9964 ($P < 0.001$) and 13.8678 ($P < 0.001$) for next 12 months prediction. The proposed framework may facilitate the study of how scholarly impact is generated, from a textual representation perspective.

1. Introduction and motivation

Estimating the scholarly impact of a scientific paper is an important and challenging problem for the academic community (Cai et al., 2019). From a micro perspective, finding and reading high-impact articles is a fundamental skill for a scholar (Gerrish and Blei, 2010). From a macro viewpoint, scholarly impact evaluation plays a crucial role in science policy and can be used in the evaluation of journals, scholars, and institutions (Wang and Barabási, 2021). For example, the scholarly impact is an important criterion in reward evaluation, funding allocation, promotion, and recruitment decisions (Cai et al., 2019; Radicchi et al., 2017; Svider et al., 2014).

Scholarly impact is commonly referenced to how scholars judge the academic influence of research (Aguinis et al., 2014). Traditionally, it is measured using the number of times (or a weighted variant of raw counts) that other scholars include a particular paper in the references section of their published work (Aguinis et al., 2014; 2012; Akella et al., 2021; Cai et al., 2019; Davis, 2008; Kaur et al., 2013; Radicchi et al., 2017). The citation-based metrics are straightforward: if more people cite an article, then more people would read it, and it is likely to have a greater impact on its research direction.

* Corresponding author.
E-mail addresses: jiangzhuoren@zju.edu.cn (Z. Jiang), lintqj@zju.edu.cn (T. Lin), huangcui@zju.edu.cn (C. Huang).

https://doi.org/10.1016/j.joi.2023.101376
Received 30 May 2022; Received in revised form 29 December 2022; Accepted 3 January 2023
Available online 27 January 2023

Although citation has already been proven effective as a scholarly impact indicator (Aksnes, 2006; Cole and Cole, 1974; Rinia et al., 1998), its practical use may be jeopardized by three major issues (Gerrish and Blei, 2010; MacRoberts and MacRoberts, 2010). **Non-Professional Documents**, such as legal documents, news articles, or blog posts, contain information that may have an impact on others, but they lack clear citations between their contents. **OCR Failures**. The historical scientific documents do contain citations, but the citation recognition is built partially based on automatic optical character recognition (OCR) technology (Cash and Hatamian, 1987), certain references may be missed because of the technology failures.[1] Besides, the improper OCR pattern may lead certain citations to incorrectly point to the wrong articles. **Newly-published Paper**. Citations are only generated after a period of time after the paper is published (Price, 1976). Therefore, there are no citation statistics for newly-published papers to measure their scholarly impact (Wang et al., 2013).

Meanwhile, since all papers are written in languages, the majority of scientific knowledge is published in text form. This inspires us to investigate the scholarly impact of a scientific paper from a textual perspective. As stated in previous studies (MacRoberts and MacRoberts, 2010), *"To determine influences on the production of a scientific article, the content of the article must be studied."* In recent years, with the development of deep learning (LeCun et al., 2015), various types of neural network models began to be widely used to solve Natural Language Processing (NLP) problems (Hirschberg and Manning, 2015). Especially, the large-scale pre-trained models (PTMs) have achieved great success and become a milestone in the research field (Brown et al., 2020; Devlin et al., 2019). In terms of language, the representation automatically learned by large-scale PTMs has been proven to capture the implicit linguistic rules and common sense hidden in the text, such as lexical meaning, syntactic structure, semantic roles, and even pragmatics (Bommasani et al., 2021; Devlin et al., 2019; Floridi and Chiriatti, 2020; Radford et al., 2021). Methods based on text representation learning have been shown to significantly improve the mining of latent semantic knowledge from scientific literature (Tshitoyan et al., 2019).

In this paper, we explore the following research questions: **RQ1**: Does the latest natural language processing technique, i.e., large-scale pre-trained model (PTM), really have a better capability for representing the content of papers? **RQ2**: How to design indicators to assess the scholarly impact of a paper based on its content representation? **RQ3**: Are the proposed indicators really positively associated with the future scholarly impact of the paper?

To address **RQ1**, we utilized a large-scale pre-trained deep model, SciBERT (Beltagy et al., 2019), for paper representation learning. SciBERT is a BERT[2]-based language model designed for performing scientific tasks. It contains 110 million parameters, and is pre-trained on a large multi-domain scientific corpus containing 1.14 million papers and 3.1 billion tokens. It has been demonstrated to make statistically significant improvements over BERT and achieve new state-of-the-art results on multiple scientific tasks. Additionally, compared with static models, such as Word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014), SciBERT, as a dynamic representation models, can not only capture word semantics, but also learn the contextually relevant information, such as word polysemy, syntactic structure, semantic roles and co-reference (Qiu et al., 2020). This characteristic has greater advantages for the representation of long texts (Peters et al., 2018). In Section 5, we validate the representation capability of SciBERT by comparing it with other representation models in multiple tasks.

Figure 1 shows an example of word representations using SciBERT and Glove. Each word is firstly encoded as a dense feature vector automatically learned by representation models; then mapped to the 2-D space using the t-SNE (Van der Maaten and Hinton, 2008) algorithm with the learned feature vector as input. SciBERT has a better capability to extract the semantic knowledge of scientific texts and embody the spatial relationships in the representation space. For instance, for the words with similar semantic roles (e.g., italy and germany, nasal and mouth) or similar lexical meanings (e.g., resumption and recovery, immunoglobulin and antibody), the representations learned by SciBERT have closer proximity compared to Glove.

To address **RQ2**, based on the paper representation learned by SciBERT, we designed two indicators: **τ (Topicality)** and **σ (Originality)**, by leveraging the spatial relations of paper representations in the semantic space to model the specific characteristics of a scientific paper. As proven in previous studies, these two characteristics can significantly affect a paper's impact. For instance, Mukherjee et al. (2017) suggested that focusing on the forefront research topics is one of the key factors for gaining future impact. Foster et al. (2015) and Wang et al. (2017) indicated the novel ideas and approaches often lead to high-impact results. In previous works, the novelty can be measured by examining whether a published paper uses combinations of referenced journals for the first time (Wang et al., 2017), detecting a published paper that consolidates existing paper clusters or connects distant ones (Chen et al., 2009; Foster et al., 2015). But these methods all require citation information for the candidate paper. In this paper, we assume the temporal spatial relations between candidate paper and high-impact papers in representation space[3] can be associated with the paper's scholarly impact. In Section 3, we provide a formal definition of the proposed indicators.

To address **RQ3**, we analyzed and discussed the relationship between proposed indicators and paper impact through regression analysis, case studies, and simulation analysis on the COVID-19 Open Research Dataset (CORD-19). Section 6 provides detailed experimental results and analysis.

[1] Based on an investigation of the citation metadata quality (Jiang et al., 2016) in the Association for Computing Machinery Digital Library, there are 18.50% of articles did not have reference metadata.

[2] BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model for deep text representation learning. It has swept in text sentiment classification, machine reading comprehension, and other 11 natural language processing tasks (Devlin et al., 2019).

[3] The temporal spatial relation is the distance or similarity relationship between a candidate paper and temporal high-impact papers in the representation space, see Section 3.2 for the detailed definition.

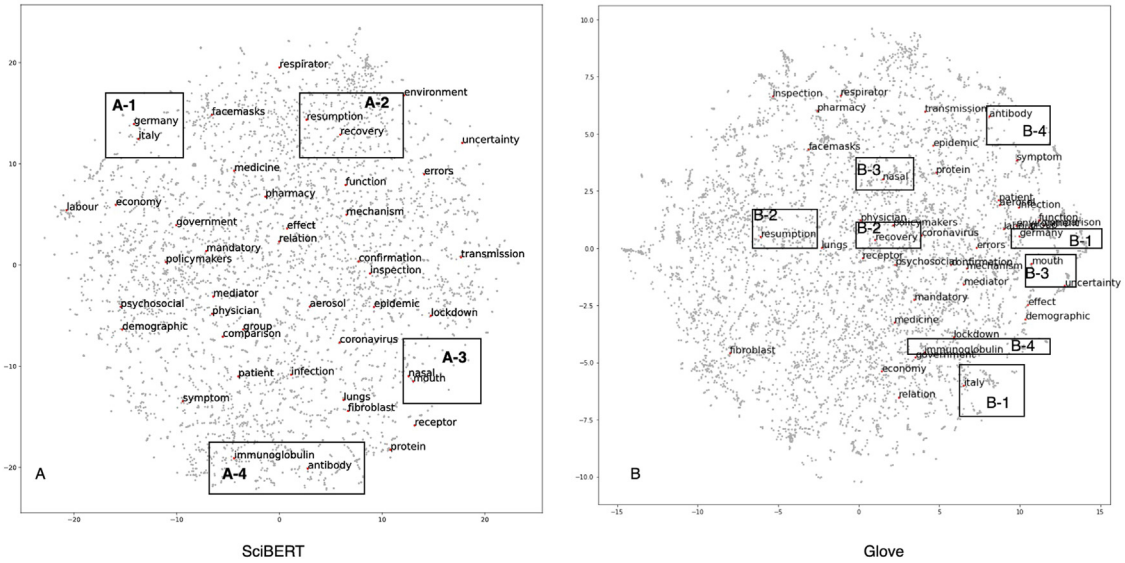

Fig. 1. The visualization of two-dimensional (2D) projection of SciBERT (A) and Glove (B) representations of 4851 noun words. The words are derived from the titles and abstracts of the papers in CORD-19 (Wang et al., 2020a). Each grey node is a word; several representative words are shown in red. The cut-out (A-1)-(B-1), (A-2)-(B-2), (A-3)-(B-3), and (A-4)-(B-4) show 4 typical word pairs in different representation spaces. The word with similar syntactic and semantic features have closer positions in SciBERT representations. The words with the top 25% of document frequency or with less than 500 document occurrences are removed for better visualization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The contribution of this paper is threefold:

- By leveraging deep representation techniques, we propose a text-content based approach that can reveal the future scholarly impact of papers.[4] Different from traditional methods, this approach is content-based and does not require human domain-specific knowledge.
- We designed two representation-based indicators to model the important paper characteristics that can be associated with the future scholarly impact. We employed multiple regression models to prove the important relationships between the proposed indicators and future impact (in next 6 months or in next 12 months).
- The simulation experiment based on regression models reproduces the process of scholarly impact generation, which may enlighten and facilitate further research on how scholarly impact is generated.

In the remainder of this paper, we (a) review relevant literature and methods, (b) introduce the adopted deep representation model, the definition of indicators, and the overall framework, (c) describe the dataset for experiments, (d) provide the detailed experimental results for representation model validations, (e) provide regression results, a case study, and the simulations based on the learned regression models, for scholarly impact analysis, and (f) discuss the contributions and limitations of our work.

2. Related work

2.1. Scholarly impact estimation

The scholarly impact of a paper is often quantified by citations and citation-based metrics (Wang and Barabási, 2021). Numerous studies consistently suggest that citations are positively correlated with assessments of scientific impact or academic recognition, such as awards, reputation (Cole and Cole, 1974), colleague evaluations (Cronin, 1996; Lawani and Bayer, 1983; Luukkonen, 1991; Oppenheim and Renn, 1978; Rinia et al., 1998), and authors' self-assessment of their scientific contributions (Aksnes, 2006).

Many studies focus on finding what factors can affect the impact of a paper (Bollen et al., 2009; Kaur et al., 2013; Radicchi et al., 2017). The pure citation-based features/indicators have been proven to have multiple limitations, such as lacking predictive power (Wang et al., 2013) and suffering the problem of missing historical metadata (Jiang et al., 2016). Besides citation-based features, various features have been explored (Bai et al., 2019), e.g., author-based features (author rank and past influence of authors, etc.) (Singh et al., 2015), journal-based features (journal prestige and journal centrality) (Bornmann et al., 2012), and paper-based features (title length and the number of references, etc.) (Haslam et al., 2008). But these features are usually sparse, and require careful design with large amounts of supplementary data collection. Furthermore, the basic information of a paper, i.e., the content

[4] The dataset and the source code: https://github.com/Lintianqianjin/Text2Impact.

of a paper, is not fully explored. Gerrish and Blei (2010) proposed a language-based approach by using dynamic topic model (Blei and Lafferty, 2006) to measuring paper's scholarly impact. Although this method is a pure text-based approach, it is a retrospective method that requires modeling the impact of a paper by computing the textual influence of a paper on subsequent papers. Therefore, the predictive ability of this method would be weak. Wang et al. (2022) found the level of innovation can significantly affect scientific literature's impact, and proposed a fine-grained text-based approach to measure the "innovation degree" of method knowledge elements in scientific literature. This motivated us to explore the scholarly impact of a paper by modeling the impact-related characteristics of the paper.

2.2. Deep representation learning

Neural models can automatically learn low-dimensional continuous vectors (distributed representation) from data as task-specific features (Bengio et al., 2013), avoiding complex feature engineering (Han et al., 2021), in contrast to previous non-neural models that mostly depended on manually-crafted features and statistical techniques. In the NLP field, static word representation models (Mikolov et al., 2013; Pennington et al., 2014), dynamic and contextualized representation models (Peters et al., 2018), especially since 2018, large-scale pre-trained language models (PTM), e.g., BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), have made a series of breakthroughs.[5] Bommasani et al. (2021) systematically explained the opportunities and risks behind large-scale pre-training models. PTMs can compensate for the shortcomings of insufficient annotated data for NLP and greatly enhance the performance of many NLP tasks. On certain datasets, the performance of PTMs can reach or even surpass human levels (Han et al., 2021). Among various PTMs, SciBERT (Beltagy et al., 2019) is specifically pre-trained and validated on a large-scale scientific paper dataset, and is shown to have a satisfactory performance on scientific tasks. Tshitoyan et al. (2019) showed that the representation models can efficiently encode materials science knowledge present in the published literature into information-dense word representations. The learned representations can benefit the material discovery. However, they only used the traditional static word representation technology which cannot learn and utilize context-related information.

Different from most previous studies, in this paper, we propose a text-based approach to exploring the scholarly impact of scientific papers without human domain-specific knowledge. By applying a large-scale pre-trained language model, i.e., SciBERT, we try to fully and effectively learn the rich semantic information from paper content and capture the important characteristics of a scientific paper that may be associated with its future impact. The following sections will introduce our methodology in detail.

3. Modeling the representation-based indicators for scholarly impact

3.1. Deep representation model

As aforementioned, in this paper, we use SciBERT (Beltagy et al., 2019) for paper representation learning. SciBERT is a pre-trained BERT-based language model for performing scientific tasks in the field of natural language processing. SciBERT is trained on papers from the corpus of semanticscholar.org. Corpus size is 1.14 million papers, 3.1 billion tokens. The full text of the papers is used for training. Similar as BERT-base model (Devlin et al., 2019), SciBERT has 12 encoder layers stacked on top of each other, which has a total of 12 attention heads and 110 million parameters.

Each paper p_i is represented by an representation vector θ_i computed by the outputs of SciBERT:

$$\theta_i = \left[\text{SciBERT}\left(\text{title}_i\right), \text{SciBERT}\left(\text{abstract}_i\right)\right] \tag{1}$$

where SciBERT$(\cdot)$ denotes the representation output from SciBERT,[6] $[\cdot,\cdot]$ is the vector concatenate operation. The paper representation is concatenated by its title representation and abstract representation. The default output dimension of SciBERT is 768, so the dimension of the final paper representation is 1536. We provide the experimental results on the validation of the representation capability of SciBERT in Section 5.

3.2. Representation-based indicator

Let us sketch the key idea. Imagine conducting a literature review, to decide whether to investigate and cite a candidate paper, a scholar may need to examine this paper from two perspectives: (1) if this candidate paper is studying a timely important topic in the field; (2) if this candidate paper is proposing a new topic or idea in the field. We summarize these two characteristics of the paper as **Topicality** and **Originality**, respectively. There are already several empirical results that have demonstrated the association between these two characteristics and scholarly impact (Fleming et al., 2007; Foster et al., 2015; Gates et al., 2019; Mukherjee et al., 2017; Wang et al., 2017; Youn et al., 2015).

In this paper, unlike the majority of previous research works, we utilize deep representation learning to model these two characteristics by computing the temporal spatial-relationship between candidate papers and seed papers in the deep representation space. Formally, we define the following notations for the proposed indicators:

[5] The different characteristics of static and dynamic representations are reviewed in (Wang et al., 2020b).

[6] The representation is generated by an average pooling for the last layer output of each token, the source code can be found at: https://github.com/Lintianqianjin/Text2Impact.

- $\phi_{i,t}$: **The impact of a candidate paper p_i at time t**. Given all the papers published before time t, the number of citations to p_i is considered as the scholarly impact of p_i. Following the previous studies (Bai et al., 2019; Cronin, 1996; Mukherjee et al., 2017; Sarigöl et al., 2014), we also assume that the scholarly impact of a paper can be reflected by the number of citations to it (Wang and Barabási, 2021).
- $P_t^\star = \{p_1^\star, \cdots, p_j^\star, \cdots\}$: **The seed paper set at time t**. Top 1% of all the papers (with $\phi_{i,\bar{t}} > 1$) sorted according to $\phi_{i,\bar{t}}$ (the number of citations from the papers published at time t). The seed papers are considered as high-impact papers (most cited papers) at time t.
- $\tau_{i,t}$: **The topicality of a paper p_i at time t**. A candidate paper p_i has a higher $\tau_{i,t}$ when the average similarity between p_i and each seed paper $p_j^\star \in P_t^\star$ is greater. High topicality (τ) means that the representation of candidate paper has a high similarity to representations of the high-impact research works (seed papers) in the field.

$$\tau_{i,t} = \frac{\sum_{p_j^\star \in P_t^\star} f(p_i, p_j^\star)}{|P_t^\star|} \tag{2}$$

where $f(\cdot,\cdot)$ is the *cosine similarity* (Xia et al., 2015) between two paper representations:

$$f(p_i, p_j) = S_{cosine}(\theta(p_i), \theta(p_j)) = \frac{\sum_{k=1}^{n} \theta_k(p_i)\theta_k(p_j)}{\sqrt{\sum_{k=1}^{n} (\theta_k(p_i))^2}\sqrt{\sum_{k=1}^{n} (\theta_k(p_j))^2}} \tag{3}$$

where $\theta(\cdot)$ is the representation function, $\theta_k(\cdot)$ indicates the k_{th} dimension of the representation.

- $\sigma_{i,t}$: **The originality of a candidate paper p_i at time t**. A paper p_i may have a higher originality $\sigma_{i,t}$ when the minimum distance of p_i from all seed papers $p_j^\star \in P_t^\star$ becomes larger.[7] High originality (σ) means that the spatial position of the candidate paper is far away from the seed papers in the representation space.

$$\sigma_{i,t} = \mathrm{Min}_{p_j^\star \in P_t^\star}(g(p_i, p_j)) \tag{4}$$

where $\mathrm{Min}(\cdot)$ is the function of taking the minimum value, $g(\cdot,\cdot)$ is the *cosine distance* between two paper representations:

$$g(p_i, p_j) = 1 - S_{cosine}(\theta(p_i), \theta(p_j)) = 1 - \frac{\sum_{k=1}^{n} \theta_k(p_i)\theta_k(p_j)}{\sqrt{\sum_{k=1}^{n} (\theta_k(p_i))^2}\sqrt{\sum_{k=1}^{n} (\theta_k(p_j))^2}} \tag{5}$$

The calculation of these two indicators (1) is time-evolving, (2) is content-based, and (3) does not require additional human domain knowledge. Please note that the proposed indicators cannot capture all the characteristics of papers that could be linked to their future impact.

3.3. Overall framework

In summary, the whole framework of this paper is shown in Fig. 2. By applying the SciBERT model on a large-scale scientific corpus, we learn the paper representations in the target paper corpus. Based on the learned representations, we obtain two temporal representation-based indicators, τ (topicality) and σ (originality), to capture two important characteristics of the candidate paper. The characteristic capture problem is then transformed into a spatial relation calculation problem in the representation space: these two indicators are based on the computing of the spatial relations between candidate papers and seed papers in the representation space. Finally, we explore the relationship between these two indicators and the future scholarly impact of the papers.

4. Dataset

4.1. Data resource

In this paper, the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020a) is utilized to validate our hypotheses. CORD-19 was first released on March 13, 2020, and has been continuously updated to date.[8] We use the version released on June 2, 2022, which contains 1,056,660 papers.

This dataset is chosen mainly for three reasons:

- As a rare contingency in human society, the studies of the COVID-19 pandemic may have little historical information (such as citations) that researchers can consult, particularly in the early stages of the outbreak. Our methodological assumptions are consistent with this practical situation.

[7] Please note that, σ only models one possible scenario of originality. We should be aware that there are other possible scenarios for modeling the originality of a paper.

[8] Historical versions of CORD-19 can be downloaded from https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html.

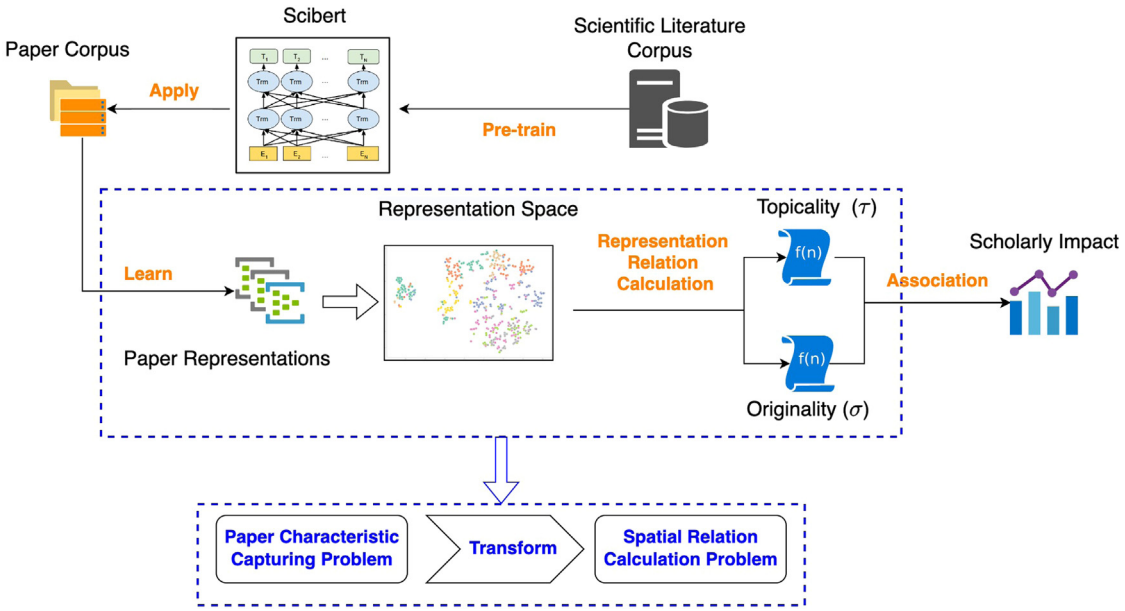

Fig. 2. The overall framework of this study.

- As a newly emerging research field, numerous new studies are produced every day, both in terms of quantity and diversity. Therefore, our approach, which can automatically analyze the information from the paper content, is suitable for COVID-19 research field.
- This COVID-19 dataset contains studies from different domains, including biology, medicine, immunology, statistics, and social sciences, etc. This makes it useful for evaluating the generalizability and applicability of our proposed approach.

4.2. Pre-processing

Paper Selection. To obtain a high-quality data set for the experiment, we adopted a series of pre-processing operations. First, we examined and removed 258,866 papers with duplicate titles. Second, we filtered and kept 234,768 papers with full text (the original dataset does not contain citation information, so we have to extract the citation information from the references in full text). Third, to obtain the papers exactly published after the COVID-19 pandemic outbreak, we further removed papers published before 2020.[9] There are 199,057 papers retained. Fourth, in the COVID-19 pandemic, scientists had posted large numbers of non-peer-reviewed papers to preprint servers, e.g., bioRxiv or medRxiv. The quality of these papers cannot be guaranteed (Kwon, 2020). To ensure the paper's quality in our experiment, we only kept papers that were published by Elsevier or indexed by Medline. Finally, 152,164 papers were retained in the experiment.

Citation Expansion. To identify the scholarly impact of papers, based on the selected 152,164 papers, we expand the paper set through citation networks. There were a total of 1,137,781 distinct papers identified through citation relations in the 152,164 papers, 260,832 of which had both title and abstract data in the CORD-19. Finally, the experimental paper dataset was composed of these 260,832 papers.

5. Validating the representation model

Before conducting experiments to explore the relationship between proposed representation-based indicators and scholarly impact, we need to validate the representation model for its capability of learning the semantic information from paper content. Specifically, we compare our adopted representation model **"SciBERT"** with two popular vector-space representation baseline models. The first baseline is a classical *sparse* text representation model **"TF-IDF"** (Aizawa, 2003), for which we assign a 21,642-d (vocabulary size, min_df = 3) vector to each paper by calculating the token in paper's title and abstract. In contrast to our method that compresses the information in the paper content into a low-dimensional dense representation, the "TF-IDF" model uses the vocabulary, and each paper is represented by its word statistic pattern. The second baseline is a *static* word embedding model **"Glove"** (Pennington et al., 2014), which is an unsupervised learning algorithm for obtaining vector representations of words. The training of "Glove" is performed on aggregated global word-word co-occurrence statistics from a corpus.

[9] The COVID-19 outbreak can be traced back to around the beginning of 2020 (Li et al., 2020).

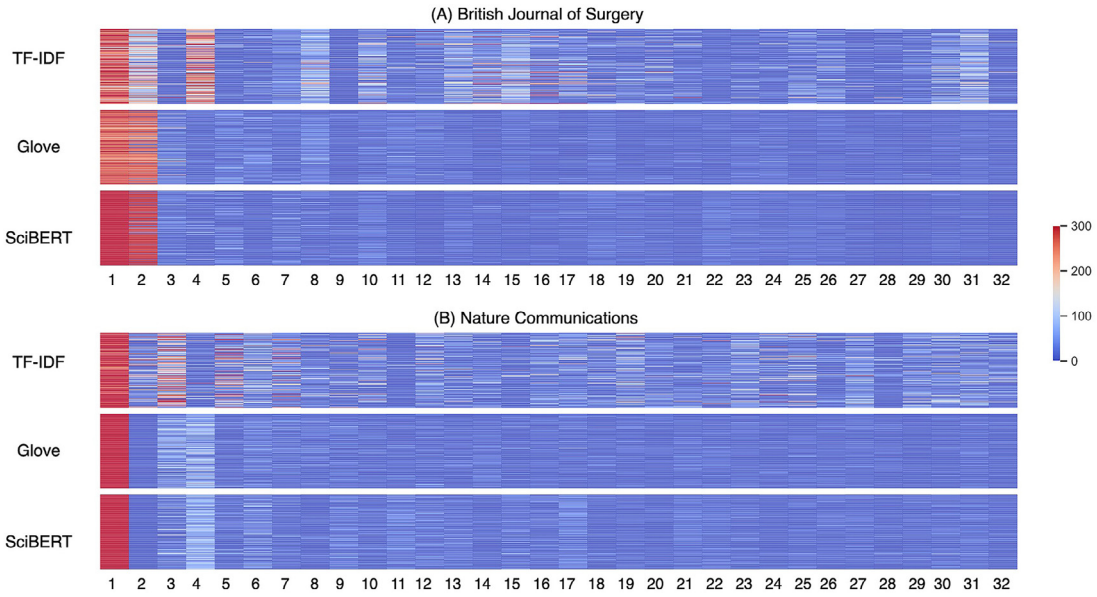

Fig. 3. Visualization (Heat map) of learned paper representations in (A) "British Journal of Surgery" and (B) "Nature Communications". Each row is a representation model, each column is a dimension, and colors depict feature values.

To ensure robustness of validation, we evaluate the employed *dense* and *dynamic* representation model against two baseline models in three tasks: (i) capturing the paper representation pattern in the journals, (ii) comparing the distributions of similarities of paper pairs with/without citation relations, and (iii) predicting the belonging journals for papers. For each task, we provide both qualitative visualization analysis and quantitative analysis with different metrics.

5.1. Comparison of paper representation pattern in the journals

Papers published in the same journal tend to have high similarities in terms of wording (Åström, 2002; Milojević, 2017) and topical semantics (D Souza and Smalheiser, 2014; Humphrey et al., 2006). Since representation model would encode the rich content information of papers into the representation space, we assume that a good representation model should enable papers in the same journal to have similar representation patterns and papers in different journals to have different representation patterns. We applied three representation models to learn the representations for papers[10] from two different journals: "British Journal of Surgery"[11] and "Nature Communications"[12]. For a fair comparison, we used Singular Value Decomposition (SVD) (Stewart, 1993) to decompose the original representation and visualize the first 32 dimensions.[13] We used a heat map to visualize the paper representation. In this heat map, each row is a representation model, each column is a dimension, and colors depict feature values. The color pattern can reflect the representation pattern. As Fig. 3 shows, from the color similarity viewpoint, it is clear that the representations (color pattern) in the same journal are similar. Compared with TF-IDF, the dense representation models (Glove and SciBERT) have significantly greater capability in capturing the consistent patterns for papers from the same journal (in many columns, the color consistency of TF-IDF is worse than the dense models). To quantitatively compare the representation patterns, we calculated standard deviations for these 32 dimensions under three representation models.[14] Generally, SciBERT has the smallest standard deviations in both journals. For instance, the averaged standard deviations of "British Journal of Surgery" with TF-IDF, Glove, and SciBERT are 73.95, 24.97, and 20.71, while the standard deviations of the first dimension of "British Journal of Surgery" with TF-IDF, Glove, and SciBERT are 114.63, 94.51, and 18.75. The averaged standard deviations of "Nature Communications" with TF-IDF, Glove, and SciBERT are 82.51, 23.93, and 23.16, while the standard deviations of the first dimension of "Nature Communications" with TF-IDF, Glove, and SciBERT are 127.71, 48.68, and 18.45. Smaller standard deviations indicate higher consistency in the paper representation patterns of journals.

Moreover, as shown in Table 1, we conducted clustering performance evaluation based on paper representations of these two journals.[15] Two clustering performance evaluation measures are reported: DaviesBouldin index (Davies and Bouldin, 1979) and

[10] In our experimental dataset, there are 197 papers from "British Journal of Surgery", 420 papers from "Nature Communications". For a fair comparison, we randomly sampled 197 papers from "Nature Communications".

[11] https://bjssjournals.onlinelibrary.wiley.com/journal/13652168.

[12] https://www.nature.com/ncomms/.

[13] The first 32 dimensions are most informative dimensions; and 32 dimensions are also clear for visualization.

[14] The detailed standard deviations for the 32 dimensions under three representation models can be found at: https://github.com/Lintianqianjin/Text2Impact.

[15] We evaluate the capability of the representation models by evaluating whether the papers can be clustered into the correct journals. If the representation model has a better capability for capturing the journal pattern, the representation learned by this model can better support the

Table 1
The DaviesBouldin Index and Silhouette Coefficient for Paper Clustering Performance Evaluation on Paper Representations from "British Journal of Surgery" and "Nature Communications".

Method	DaviesBouldin Index ↓	Silhouette Coefficient ↑
TF-IDF	7.55	0.003
Glove	4.75	0.013
SciBERT	**4.41**	**0.033**

Bold font indicates the best results. ↑ indicates the greater the better, ↓ indicates the smaller the better.

Table 2
The Jensen-Shannon Divergence and Shannon Entropy of Similarity Distributions of Paper Pairs with/without Citation Relations.

Method	Divergence	Entropy	
		with Citation	without Citation
TF-IDF	0.2244	3.7837	2.8067
Glove	0.0428	3.2731	3.3519
SciBERT	0.1054	3.4402	3.4712

Silhouette coefficient (Rousseeuw, 1987). SciBERT achieves the best clustering performance on both evaluation measures. These observations indicate SciBERT has the best capability for capturing the journal pattern among the three models.

5.2. Distributions of distance/similarity of paper pairs with/without citation relations

To further validate the learning capability of the representation model for scientific papers, we compared the distributions of the distance/similarity of paper pairs with/without citation relations. We assumed that (1) the distance/similarity of paper pairs with citation relations and paper pairs without citation relations should be significantly different. (2) The representation similarity distribution of paper pairs without citation relations should be more random (with higher uncertainty) than paper pairs with citation relations. These two qualities should be accurately reflected by a good representation model.

There are 779,314 pairs of papers with citation relations in the experiment dataset. For comparison, we randomly sampled 779,314 paper pairs without citation relations.[16] Then we applied three representation models to learn the representations for the paper pairs. From the difference viewpoint, as shown in Fig. 4, compared to Glove, the similarity/distance distribution pattern (shape and position of red and blue areas) of SciBERT and TF-IDF show more significant differences. We further calculated the Jensen-Shannon divergence (Lin, 1991) to quantitatively compare the difference between distributions with/without citation relations. As shown in Table 2, the divergences of SciBERT (0.1054) and TF-IDF (0.2244) are greater than Glove (0.0428). To explicitly compare the uncertainty of different types of paper pairs, we calculated the Shannon entropy (Shannon, 1948) of different types of paper pairs under different representations.[17] As shown in Table 2, with TF-IDF representation, the entropy of paper pairs without citation relation (2.8067) is much smaller than the entropy of paper pairs with citation relation (3.7837). In contrast, with the Glove and SciBERT representations, the entropy of paper pairs without citation relation is slightly greater than the entropy of paper pairs with citation relation. A possible explanation is that since TF-IDF is a one-hot representation (Turian et al., 2010), the similarity can only be calculated if the papers use exactly the same words. For the papers that are semantically similar but use different words, the similarity will be zero. This problem may reduce the performance of the TF-IDF representations. Overall, SciBERT performs better in terms of both difference and uncertainty: the representation learned by SciBERT not only effectively distinguishes whether there are citation relations between them, but also provides a superior semantic representation capability to represent papers.

5.3. Predicting the belonging journal for papers

Motivated by previous works (Peng et al., 2021; Wang et al., 2016; Zhang et al., 2019), we conducted the journal paper visualization and journal classification/clustering task to further validate the representational capability of the representation model.

As shown in Fig. 5, we visualized the paper representations from the three models. For better visualization, we randomly sampled 50 papers from each of the 12 journals.[18] In terms of node spatial position, the TF-IDF representation cannot distinguish different journals, as paper nodes of different colors are mixed together. Meanwhile, SciBERT has a better distinguishing capability than Glove:

clustering algorithm to automatically assign the papers to the correct journals. All papers from "British Journal of Surgery" (197) and "Nature Communications" (420) are used for clustering evaluation.

[16] We examined the randomly sampled paper pairs and removed the self-citation pairs and the pairs with citation relations. Finally, we got 777,255 paper pairs without citation relations.

[17] The source code of calculations of Jensen-Shannon divergence and Shannon entropy can be found at https://github.com/Lintianqianjin/Text2Impact.

[18] We try to select journals in diverse fields, such as "Morbidity and Mortality Weekly Report" (MMWR Morb Mortal Wkly Rep), "Foods" (Foods), and "British Journal of Surgery" (Br J Surg). Detailed visualized journal list can be found at: https://github.com/Lintianqianjin/Text2Impact.

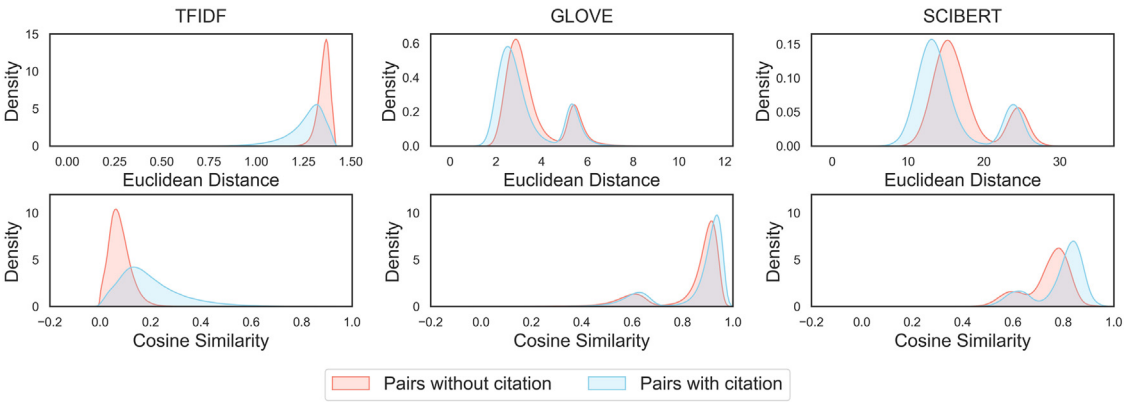

Fig. 4. The visualization of the distributions of paper pairs with/without citation relations. Each column illustrates a representation model. The x-axis represents the Euclidean distance (the upper row) or cosine similarity (the lower row) of the paper pairs, and the y-axis represents the probability density. Blue area indicates pairs of papers with citation relations; red area indicates randomly selected papers pairs (without citation relations). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. The visualization of papers from 12 journals. Each node is a paper, its color denotes the belonging journals. Each paper is firstly encoded as feature vectors learned by TF-IDF, Glove, and SciBERT; then mapped into the 2-D space using the t-SNE (Van der Maaten and Hinton, 2008) algorithm with learned feature vector as input. For better visualization, we randomly sampled 50 papers from each of the 12 journals.

the positions of nodes of the same color are more aggregated, and the distinction between the positions of nodes of different colors is clearer.

To quantitatively compare the representation models, we performed 49 sets of clustering performance evaluations based on paper representations of journals containing the most papers (from top 2 to top 50 journals) in the experimental dataset.[19] As shown in Fig. 6, in almost all settings, SciBERT consistently achieves optimal results on both evaluation metrics. These experimental results show that SciBERT has the best journal differentiation capability and the best capability to represent paper content, among the three models.

Moreover, we conducted a classification task on predicting the belonging journals based on the different representation models. To comprehensively validate the predicting capability of representation models, we designed two experimental settings. (1) **Balanced Classification**, we selected 6 journals with different research areas for classification. As different journals contain different numbers of papers, we randomly selected 197 papers[20] in each journal (totally 1182 papers). (2) **Imbalanced Classification**, we choose all the papers from 10 journals containing the most papers (totally 20,897 papers) for classification.[21] Meanwhile, as this study is conducted on the COVID-19 Open Research Dataset, BioBERT (Lee et al., 2020) trained specifically on biomedical dataset may achieve good representation performance. For this task, we also included BioBERT for comparison. A random forest classifier (Breiman, 2001) was utilized for all representation models.[22] 5-fold cross validation was applied for avoiding the data distribution bias. The results are shown in Table 3. By using the representation learned by SciBERT, the classification model achieved the best performance with both

[19] The journal list can be found at: https://github.com/Lintianqianjin/Text2Impact. There are totally 42,154 papers used for evaluation.

[20] The 6 journals are “International Journal of Environmental Research and Public Health” (Int J Environ Res Public Health), “Frontiers in Public Health” (Front Public Health), “British Journal of Surgery” (Br J Surg), “Frontiers in Immunology” (Front Immunol), “International Journal of Molecular Sciences” (Int J Mol Sci), and “Journal of Clinical Medicine”(J Clin Med). “British Journal of Surgery” contains the fewest number of papers (only 197); to ensure that each journal contains the same number of papers, we used all papers from “British Journal of Surgery” and randomly sampled 197 papers from the other 5 journals.

[21] The number of papers contained in different journals varies greatly, for example, “International Journal of Environmental Research and Public Health” has 5242 papers and “Journal of Clinical Medicine” has only 1291 papers.

[22] The code can be found at: https://github.com/Lintianqianjin/Text2Impact.

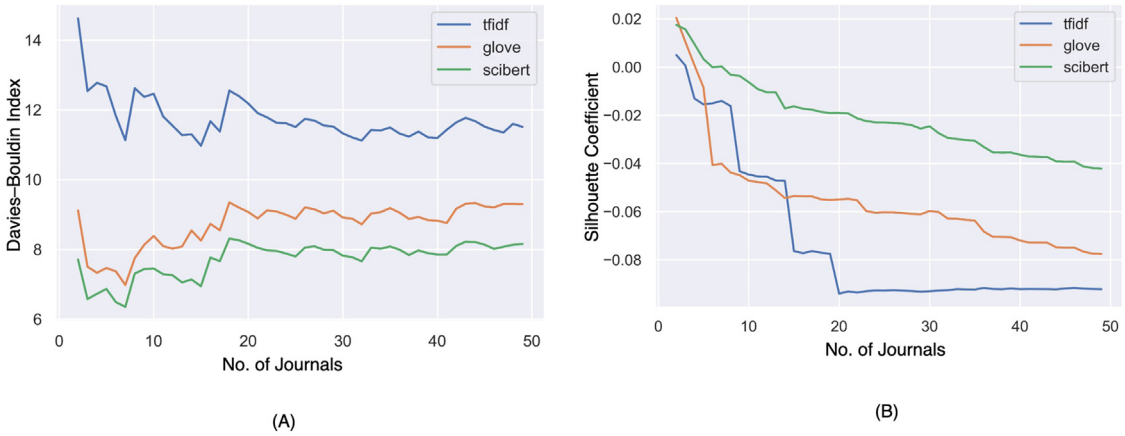

Fig. 6. (A) The trends of DaviesBouldin index (the smaller the better) and (B) the trends of Silhouette coefficient (the greater the better) of clustering performance based on the paper representations of journals containing the most papers (from top 2 to top 50 journals).

Table 3
Classification Performance Based on Different Representation Models.

Method	Balanced Classification		Imbalanced Classification	
	Macro F1	Micro F1	Macro F1	Micro F1
TF-IDF	0.6421 (±0.0333)	0.6506 (±0.0314)	0.4393 (±0.0076)	0.5129 (±0.0048)
Glove	0.6592 (±0.0317)	0.6632 (±0.0320)	0.4491 (±0.0031)	0.4978 (±0.0075)
BioBERT	0.6855 (±0.0257)	0.6887 (±0.0245)	0.4923 (±0.0082)	0.5400 (±0.0022)
SciBERT	**0.6878** (±0.0363) **(7.1%↑, 3.9%↑)**	**0.6921** (±0.0350) **(6.4%↑, 3.7%↑)**	**0.4999** (±0.0077) **(13.8%↑, 8.6%↑)**	**0.5452** (±0.0066) **(9.5%↑, 5.6%↑)**

Format of result: Mean Value (± Standard Deviation). Bold font indicates the best results. ↑ indicates maximum improvement, ↑ indicates average improvement.

evaluation metrics on both classification tasks. Please note that, the main purpose of this experiment is not to pursue the optimal performance of journal classification, but to compare the content semantic learning capability of different representation models. For the same classification model, the performance difference based on the different representation models can demonstrate the capability differences in text representation learning.

To conclude, by analyzing and validating representation models in three tasks, we demonstrate the advantages of SciBert over traditional representation models.

6. The relationship between representation-based indicator and scholarly impact

6.1. Regression analysis

The linear regression analysis (Montgomery et al., 2021) was performed to examine the relationship between proposed representation-based indicators and paper scholarly impact. The regression model would estimate the impact $\phi_{i,t}$[23] of a random paper p_i at time t. Specifically, the linear models were used to model the dependence of the paper impact ϕ on features τ (topicality) and σ (originality).[24] The learned relationships were linear and can be written as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i xi \tag{6}$$

The predicted outcome of a paper instance was a weighted sum of its i_{th} features. The β_j represented the learned feature coefficients. β_0 was the intercept. The τ and σ were treated as independent variables (features). In order to validate our proposed indicator as thoroughly as possible, we chose two representative time windows: 6 months (representing the short to medium term), and 12 months (representing the medium to long term). These settings of time windows have been frequently used in previous research (Breitzman, 2021; Chakraborty et al., 2014; Croft and Sack, 2022; Eysenbach, 2006). Therefore, there were two types of dependent variable ϕ to be estimated. $\phi_{i,6}$, the impact of a candidate paper p_i in next 6 months, and $\phi_{i,12}$, the impact of a candidate paper p_i in next 12

[23] In this paper, as defined in Section 3.2, $\phi_{i,t}$ is the citation count of p_i at time t.

[24] To avoid possible statistical issues and make sure the indicators τ (topicality) and σ (originality) are not oppositely correlated, we calculated the rank correlation coefficient between the paper rankings based on these two indicators. The Spearman rank correlation coefficient is 0.29 ($p < 0.001$), Kendall rank correlation coefficient is 0.27 ($p < 0.001$). τ and σ show a very weak correlation.

Table 4
Regression results for three models that estimate the paper scholarly impact in next 6 months.

	coef	std err	t	**P**> |t|	[0.025	0.975]
const	-0.4025	0.6440	-0.6250	0.5320	-1.6640	0.8590
σ	3.2935***	0.8820	3.7350	0.0000	1.5650	5.0220
	coef	**std err**	**t**	**P**> |t|	**[0.025**	**0.975]**
const	1.3314	0.1120	11.8840	0.0000	1.1120	1.5510
τ	5.1629***	0.8090	6.3780	0.0000	3.5760	6.7500
	coef	**std err**	**t**	**P**> |t|	**[0.025**	**0.975]**
const	-2.8659	0.7140	-4.0160	0.0000	-4.2650	-1.4670
σ	5.4915***	0.9220	5.9550	0.0000	3.6840	7.2990
τ	6.6879***	0.8480	7.8890	0.0000	5.0260	8.3500

Levels of statistical significance: $P < 0.05$ (∗), $P < 0.01$ (∗∗), $P < 0.001$ (∗∗∗).

Table 5
Regression results for three models that estimate the paper scholarly impact in next 12 months.

	coef	std err	t	**P**> |**t**|	[0.025	0.975]
const	-2.9870	1.1950	-2.5000	0.0120	-5.3290	-0.6450
σ	7.7893***	1.6710	4.6630	0.0000	4.5140	11.0640
	coef	**std err**	**t**	**P**> |**t**|	**[0.025**	**0.975]**
const	1.0381	0.2090	4.9620	0.0000	0.6280	1.4480
τ	10.6447***	1.3540	7.8600	0.0000	7.9900	13.3000
	coef	**std err**	**t**	**P**> |**t**|	**[0.025**	**0.975]**
const	-8.7008	1.3230	-6.5750	0.0000	-11.2950	-6.1070
σ	12.9964***	1.7440	7.4520	0.0000	9.5780	16.4150
τ	13.8678***	1.4170	9.7850	0.0000	11.0900	16.6460

Levels of statistical significance: $P < 0.05$ (∗), $P < 0.01$ (∗∗), $P < 0.001$ (∗∗∗).

months. There are 10,304 paper utilized for the prediction of next 6 months' impact, and there are 7970 papers for the prediction of next 12 months' impact. The Ordinary Least Squares (OLS) method (Hutcheson, 2011) was used to find the weights that minimize the squared differences between the actual and the estimated outcomes.[25]

To ensure the robustness of regression analysis, we performed regression learning not only for each independent variable separately (bivariate regression analysis), but also for both independent variables simultaneously (multivariate regression analysis). Therefore, there were totally 6 sets of regression modeling conducted. The regression coefficients, standard errors of estimation, absolute values of the t-statistic, significance levels, and 95% confidence intervals of these regression models are reported in Tables 4 and 5.

The following observations can be made:

(1) For the impact of a paper in 6 months after publication (Table 4), both σ and τ are significantly positive for all the regression models, i.e., for the bivariate regression analysis of σ, the coefficient is 3.2935, 95% CI[1.5650,5.0220] ($P < 0.001$) ; for the bivariate regression analysis of τ, the coefficient is 5.1629, 95% CI[3.5760,6.7500] ($P < 0.001$); for the multivariate regression analysis, the coefficient of σ is 5.4915, 95% CI[3.6840,7.2990] ($P < 0.001$), while the coefficient of τ is 6.6879, 95% CI[5.0260,8.3500] ($P < 0.001$). Overall, in the short to medium term, a paper with high σ and τ would have potential to achieve high impact. Relatively, τ has a greater effect on the impact of papers than σ.

(2) For the impact of a paper in 12 months after publication (Table 5), both σ and τ are still significantly positive for all the regression models, i.e., for the bivariate regression analysis of σ, the coefficient is 7.7893, 95% CI[4.5140,11.0640] ($P < 0.001$) ; for the bivariate regression analysis of τ, the coefficient is 10.6447, 95% CI[7.9900,13.3000] ($P < 0.001$); for the multivariate regression analysis, the coefficient of σ is 12.9964, 95% CI[9.5780,16.4150] ($P < 0.001$), while the coefficient of τ is 13.8678, 95% CI[11.0900,16.6460] ($P < 0.001$). Overall, in the medium to long term, a paper with high σ and τ still has potential to achieve high impact. τ has a relatively greater effect on the impact of papers than σ; but compared with 6 months time-window, the difference in the coefficients of τ and σ on the scholarly impact is smaller.

6.2. Case study

In Fig. 7, we provide a representative case in the papers published in January and February. The red triangle indicates a seed paper in February (Top 1% cited paper in February) which is entitled *"A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster"*. Two blue triangles indicate two high-impact papers in March (Top 1% cited paper in March). Paper A entitled *"Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia*

[25] We strictly control the citation time-window. For each paper, its impact is only obtained from papers published in the predicted time range. The code of regression analysis can be found at https://github.com/Lintianqianjin/Text2Impact.

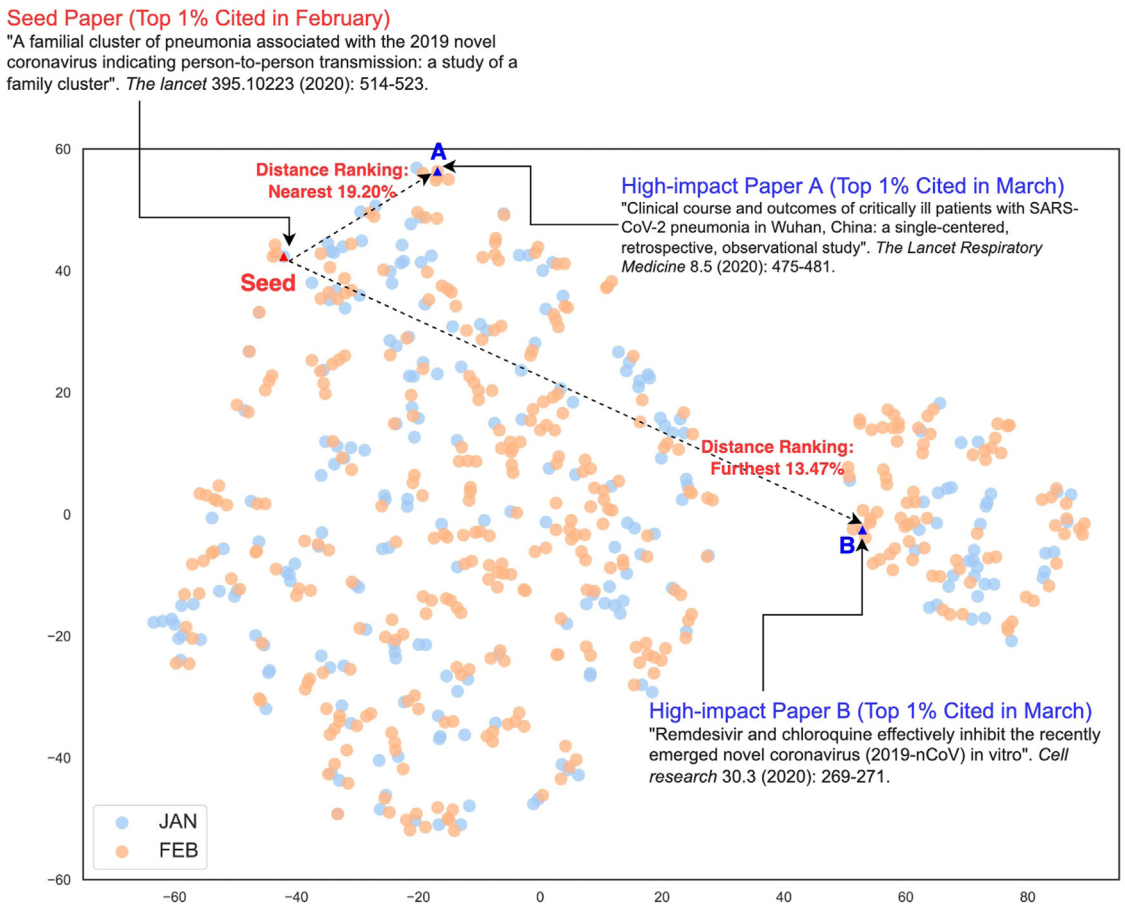

Fig. 7. The visualization of two-dimensional (2D) projection of SciBERT representations of all papers published in January and February, 2020; t-SNE algorithm is used for dimensionality reduction. The blue nodes are papers published in January and the orange nodes are papers published in February. The red triangle indicates a seed paper in February (Top 1% cited paper in February). The blue triangles indicate two typical high-impact papers in March (Top 1% cited paper in March): paper A is close to seed paper (top 19.20% nearest in Euclidean distance ranking), and paper B is far from seed paper (top 13.47% furthest in Euclidean distance ranking). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Wuhan, China: a single-centered, retrospective, observational study" is in the top 19.20% of papers nearest to seed paper. Generally, both papers are case studies of COVID-19 patients, which was a popular research topic in early 2020 (the beginning of the COVID-19 outbreak). Therefore, paper A had a high topicality with the seed paper. Correspondingly, the spatial distance between paper A and seed paper is close in representation space, which is consistent with the hypothesis of the proposed indicator τ (topicality). Meanwhile, paper B entitled *"Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro"* is in the top 13.47% of papers far from seed papers. Paper B studies a different topic with seed paper, which is focusing on antiviral treatment. Therefore, paper B has a high originality compared to the seed paper. Correspondingly, the spatial distance between paper B and seed paper is large in representation space, which is consistent with the hypothesis of the proposed indicator σ (originality). In summary, this case demonstrates that the proposed indicators, τ and σ, can capture the corresponding characteristics of the paper by computing the spatial relationships between candidate paper and seed papers in the representation space.

6.3. Simulation analysis

Based on the learned multivariate regression models (for estimating the scholarly impact in next 6 months and next 12 months), we simulated the potential impacts of papers in a two-dimensional representation space with different settings of the seed paper(s). Specifically, first, the positions of seed papers were randomly generated in the representation space according to a uniform distribution. Second, for a spatial position of a candidate paper, its τ and σ can be calculated based on its spatial relations with seed paper(s) via Eqs. (2) and (4). Third, the learned regression model can predict the impact of the candidate paper based on its τ and σ. As Fig. 8 shows, red triangles are seed papers, areas with deeper grey color indicate higher chance to be high-impact (get more citations).

The following observations can be made:

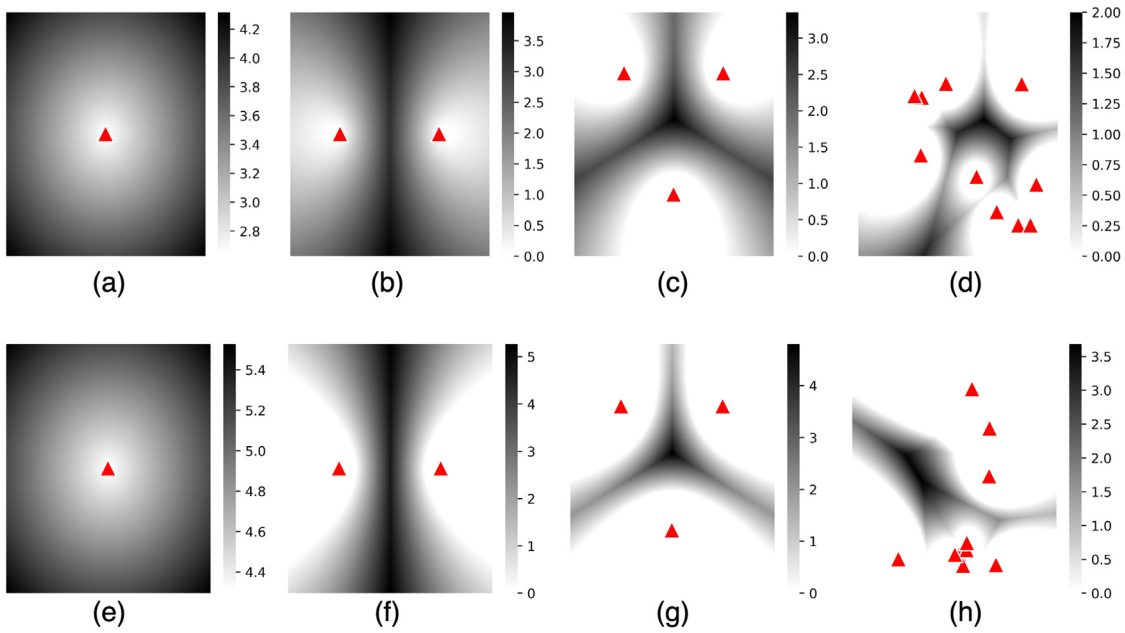

Fig. 8. Simulations based on the learned multivariate regression models. Sub-plot (a), (b), (c), and (d) show the simulation of the potential impact (the possible number of citations) of papers at different spatial positions in the representation space for the next 6 months, with one seed paper, two seed papers, three seed papers, and ten seed papers, respectively. Sub-plot (e), (f), (g), and (h) show the simulation of the potential impact of papers at different spatial positions in the representation space for the next 12 months, with one seed paper, two seed papers, three seed papers, and ten seed papers, respectively. The red triangles indicate the spatial positions of seed papers, the area with deeper color indicates that if a paper's representation is located in this area, it has a higher chance of being high-impact.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- Simulations based on the multivariate regression models of the next 6 months and the next 12 months both show a similar pattern.
- The areas very close to the seed papers (around the red triangles) are almost white. This phenomenon suggests that a candidate paper will have difficulty in achieving high impact if its representation is too similar to that of a seed paper.
- For the case of only one seed paper, the darker a spatial position is, the further it is from the seed paper. According to this observation, a candidate paper with higher σ (the representation of candidate papers is far away from that of seed paper) has a better probability of achieving a high impact.
- The cases of two seed papers and three seed papers are similar. If a spatial position is located between the seed papers or far from the seed papers in the representation space, its color is darker. Therefore, if a candidate paper has a high similarity to all seed papers (with high τ), it has a greater chance of achieving high impact; if the distances between the candidate paper and seed papers are all large (with high σ), the candidate paper is possible to obtain a high impact in the future.
- For the case of ten seed papers, if a spatial position is located in a narrow intermediate region surrounded by multiple seed papers, its color is darker. Meanwhile, there are some narrow areas away from all the seed papers which are darker in color. This indicates the papers that can cover multiple popular research topics could be possible high-impact papers; while original papers still have a chance to obtain a high scholarly impact.

We can explain the observed phenomenon from two perspectives:

- From the perspective of *knowledge creation*: (1) new ideas often synthesize existing knowledge. For example, based on the analysis of U.S. patents, new inventions are often made by recombining existing technologies (Youn et al., 2015). Evidence obtained in a wide range of surveys consistently suggests that uncommon combinations in scientific publications usually mean that the paper has a higher probability of achieving high impact (Wang and Barabási, 2021). Thus, in the representation space, papers that are located in the region between multiple seed papers may integrate multiple research ideas, thus can achieve high impact in the future. (2) Proposing atypical research ideas that are completely different from the current research works, e.g., paradigm-changing discoveries, is one of the ways to achieve high impact (Wang et al., 2017). This explains why papers that are far from all seed papers in the representation space always have a chance to be high-impact papers.
- From the perspective of *disciplinary development*: (1) when the research field is in its early stages (only a few seed papers are available), there is enormous room for becoming a high-impact paper in the future (the dark area is large). As the research field continues to become mature, i.e., there are more and more seed papers, then there is less room for possible high-impact papers (the dark area is getting smaller). (2) if we treat different seed papers as the core papers in different disciplines, these intermediate areas surrounded by seed papers can be considered as the representation areas of interdisciplinary research, and these areas are

darker in color (higher chance to be high-impact). Therefore, interdisciplinary research (Gates et al., 2019), by combining the findings from different fields, has the potential to produce high-impact research works (Larivière et al., 2015; Sinatra et al., 2015).

7. Conclusions and future works

7.1. Conclusions

Based on the deep textual representation of scientific papers, this paper proposes two indicators for revealing the future scholarly impact of papers. We chose the COVID-19 Open Research Dataset for experimental validation and analysis. This dataset was chosen to better validate the generalization and applicability of the proposed approach due to the substantial number of COVID-19-related papers and the diversity of research domains they cover. Besides, in the early stages of the COVID-19 outbreak, there may be little historical information (such as citations) for scholars to consult. This practical situation is consistent with our methodological assumptions. Through comprehensive experiments, we validate the effectiveness of the employed large-scale representation model. The regression experiment results indicate that the proposed indicators are positively and significantly associated with a paper's future scholarly impact. The simulation analysis and the case studies further demonstrate the soundness of our proposed indicators.

The findings from this study can be important for scientific bibliometrics and academic retrieval/recommendation. Although several citation-based features and indicators have been proposed in the past, few studies have addressed a pure text-based indicator. This study achieved this by utilizing the state-of-the-art deep representation model (SciBERT). The comprehensive validation of SciBERT in this study can provide empirical evidence for its superior semantic representation capability and great potential in future academic evaluation studies. For the future use of the proposed indicators, we suggest that: (1) they can be used directly to assist academic evaluation. Since the proposed indicators only require the textual content of the paper, they are friendly to the newly published papers without citation information. (2) They can be used as features for constructing complex academic evaluation or prediction models. (3) Simulations can be done with these two indicators to further explore the formation mechanism of scholarly impact. In order to help other scholars to reproduce the experiment outcome and use the proposed indicators for further research, we release the code via https://github.com/Lintianqianjin/Text2Impact. Please note that, firstly, the representation-based indicators are not proposed to replace the current ones, but rather to provide a useful supplement and an interesting analytical perspective. Secondly, although we use highly-cited papers as seed papers to calculate indicators, other methods can be used to define seed papers, and different methods of defining seed papers may result in various indicator values for the same candidate paper.

7.2. Limitations and future works

This study has three potential limitations: (1) All citation information is generated from papers in the experimental dataset. If the external citation information can be added, the quality of the dataset for experimental analysis can be further improved. (2) In this study, we used one dataset for experimental analysis. Our results and analysis would be more robust if we could validate the indicators on various datasets. (3) The proposed indicators cannot capture all impact-related characteristics of a paper.

In the future, we will further optimize the indicator design to make them more compatible with the corresponding characteristics. Meanwhile, to validate the universality of the proposed approach, we will investigate this problem by using more datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (72104212), the State Key Program of National Natural Science of China (72134007), the Natural Science Foundation of Zhejiang Province (LY22G030002), the Central Government Guiding Project for the Development of Local Science and Technology of Zhejiang Province (2021ZY1004), and the Fundamental Research Funds for the Central Universities. The model training and prediction are supported by Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. We thank Weikang Yuan for his contribution to the supplementary experiments during the revision of the paper. We appreciate the editors' and anonymous reviewers' helpful and insightful comments.

References

Aguinis, H., Shapiro, D. L., Antonacopoulou, E. P., & Cummings, T. G. (2014). Scholarly impact: A pluralist conceptualization. *Academy of Management Learning & Education, 13*(4), 623–639.

Aguinis, H., Suárez-González, I., Lannelongue, G., & Joo, H. (2012). Scholarly impact revisited. *Academy of Management Perspectives, 26*(2), 105–132.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45–65.

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*(2), 101128.

Aksnes, D. W. (2006). Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology, 57*(2), 169–185.

Åström, F. (2002). Visualizing library and information science concept spaces through keyword and citation based maps and clusters. In *Emerging frameworks and methods: Proceedings of the fourth international conference on conceptions of library and information science (COLIS4)* (pp. 185–197). Libraries Unlimited Greenwood Village.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*(1), 407–418.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3615–3620).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one, 4*(6), e6022.
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics, 6*(1), 11–18.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Breitzman, A. (2021). The relationship between web usage and citation statistics for electronics and information technology articles. *Scientometrics, 126*(3), 2085–2105.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Cai, L., Tian, J., Liu, J., Bai, X., Lee, I., Kong, X., & Xia, F. (2019). Scholarly impact assessment: A survey of citation weighting solutions. *Scientometrics, 118*(2), 453–478.
Cash, G. L., & Hatamian, M. (1987). Optical character recognition by the method of moments. *Computer Vision, Graphics, and Image Processing, 39*(3), 291–310.
Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM joint conference on digital libraries* (pp. 351–360). IEEE.
Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*(3), 191–209.
Cole, J. R., & Cole, S. (1974). Social stratification in science. *American Journal of Physics, 42*(10), 923–924.
Croft, W. L., & Sack, J.-R. (2022). Predicting the citation count and citescore of journals one year in advance. *Journal of Informetrics, 16*(4), 101349.
Cronin, B. (1996). Research brief rates of return to citation. *Journal of Documentation*.
D Souza, J. L., & Smalheiser, N. R. (2014). Three journal similarity metrics and their application to biomedical journals. *PloS one, 9*(12), e115681.
Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224–227.
Davis, P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology, 59*(13), 2186–2188.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).
Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology, 4*(5), e157.
Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly, 52*(3), 443–475.
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908.
Gates, A. J., Ke, Q., Varol, O., & Barabási, A.-L. (2019). Nature's reach: Narrow work has broad impact.
Gerrish, S. M., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 375–382).
Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al., (2021). Pre-trained models: Past, present and future. *AI Open, 2*, 225–250.
Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.
Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.
Humphrey, S. M., Lu, C. J., Rogers, W. J., & Browne, A. C. (2006). Journal descriptor indexing tool for categorizing text according to discipline or semantic type. In *AMIA annual symposium proceedings: vol. 2006* (p. 960). American Medical Informatics Association.
Hutcheson, G. D. (2011). Ordinary least-squares regression. In L. Moutinho, & G. D. Hutcheson (Eds.), *The SAGE dictionary of quantitative management research* (pp. 224–228).
Jiang, Z., Liu, X., & Chen, Y. (2016). Recovering uncaptured citations in a scholarly network: A two-step citation analysis to estimate publication importance. *Journal of the Association for Information Science and Technology, 67*(7), 1722–1735.
Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics, 7*(4), 924–932.
Kwon, D. (2020). How swamped preprint servers are blocking bad coronavirus research. *Nature, 581*(7807), 130–132.
Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *PloS one, 10*(3), e0122565.
Lawani, S. M., & Bayer, A. E. (1983). Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment. *Journal of the American Society for Information science, 34*(1), 59–66.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.
Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... Wong, J. Y., et al., (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.
Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory, 37*(1), 145–151.
Luukkonen, T. (1991). Citation indicators and peer review: Their time-scales, criteria of evaluation, and biases. *Research Evaluation, 1*(1), 21–30.
Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11).
MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology, 61*(1), 1–12.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*.
Milojević, S. (2017). The length and semantic structure of article titles-evolving disciplinary practices and correlations with impact. *Frontiers in Research Metrics and Analytics, 2*, 2.
Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances, 3*(4), e1601315.
Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science, 29*(5), 225–231.
Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances, 7*(17), eabb9004.
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. 10.18653/v1/N18-1202. https://aclanthology.org/N18-1202
Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science, 27*(5), 292–306.
Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences, 63*(10), 1872–1897.
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Radicchi, F., Weissman, A., & Bollen, J. (2017). Quantifying perceived impact of scientific publications. *Journal of Informetrics, 11*(3), 704–712.
Rinia, E. J., Van Leeuwen, T. N., Van Vuren, H. G., & Van Raan, A. F. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the netherlands. *Research Policy, 27*(1), 95–107.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science, 3*, 1–16.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.
Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A.-L. (2015). A century of physics. *Nature Physics, 11*(10), 791–796.
Singh, M., Patidar, V., Kumar, S., Chakraborty, T., Mukherjee, A., & Goyal, P. (2015). The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1271–1280).
Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review, 35*(4), 551–566.
Svider, P. F., Lopez, S. A., Husain, Q., Bhagat, N., Eloy, J. A., & Langer, P. D. (2014). The association between scholarly impact and national institutes of health funding in ophthalmology. *Ophthalmology, 121*(1), 423–428.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98.
Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384–394).
Wang, D., & Barabási, A.-L. (2021). *The science of science*. Cambridge University Press.
Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436.
Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R. M., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D. M., Weld, D. S., Etzioni, O., & Kohlmeier, S. (2020a). CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. https://aclanthology.org/2020.nlpcovid19-acl.1
Wang, S., Tang, J., Aggarwal, C., & Liu, H. (2016). Linked document embedding for classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 115–124).
Wang, Y., Hou, Y., Che, W., & Liu, T. (2020b). From static to dynamic word representations: A survey. *International Journal of Machine Learning and Cybernetics, 11*(7), 1611–1630.
Wang, Z., Wang, K., Liu, J., Huang, J., & Chen, H. (2022). Measuring the innovation of method knowledge elements in scientific literature. *Scientometrics, 127*(5), 2803–2827.
Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences, 307*, 39–52.
Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: Evidence from us patents. *Journal of the Royal Society interface, 12*(106), 20150272.
Zhang, Y., Zhao, F., & Lu, J. (2019). P2V: Large-scale academic paper embedding. *Scientometrics, 121*(1), 399–432.