Scientometrics (2022) 127:4315–4333
https://doi.org/10.1007/s11192-022-04462-y

# Cover papers of top journals are reliable source for emerging topics detection: a machine learning based prediction framework

**Wenjie Wei[1,2] · Hongxu Liu[1] · Zhuanlan Sun[3]**

Received: 19 April 2021 / Accepted: 30 June 2022 / Published online: 18 July 2022


**Abstract**
The detection of emerging trends is of great interest to many stakeholders such as government and industry. Previous research focused on the machine learning, network analysis and time series analysis based on the bibliometrics data and made a promising progress. However, these approaches inevitably have time delay problems. For the reason that leader papers of "emerging topics" share the similar characters with the "cover papers", this study present a novel approach to translate the "emerging topics" detection to "cover paper" prediction. By using "AdaBoost model" and topic model, we construct a machine learning framework to imitate the top journal (chief) editor's judgement to select cover paper from material science. The results of our prediction were validated by consulting with field experts. This approach was also suitable for the Nature, Science, and Cell journals.

**Keywords** Cover paper · Emerging topics detection · Research trends prediction · Machine learning · Text mining · Topic model

## Introduction

The evolution of research topics can be distinguished in three main stages: (i) embryonic stage, (ii) early stage, and (iii) recognized stage (Salatino, 2019). Prediction of future research trends has been a long-standing need for many stakeholders: including research evaluation, funding or awarding research policy making, and even commercial investment. Traditionally, expert interviews, workshops with experts and first-line stakeholders, and Delphi studies were used to identify the current research trends (Rotolo et al., 2015). However, a vast and rapidly growing literature makes it more and more difficult to identify the scientific research trends than ever before (Gibney, 2014). Manually identify emerging

✉ Zhuanlan Sun
zlsuen@163.com

[1] Tongji University Library, Tongji University, Shanghai 200092, China

[2] College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China

[3] Institute of High-Quality Development Evaluation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

topics that will have long-term scientific impact is becoming time-consuming and resource-intensive, especially in their early stage of post-publication. Because the evaluation of whether an idea is novel or surprising depends crucially on already-existing knowledge, however, we are facing the literature explosion challenge (Bornmann & Mutz, 2015).

To overcome this problem, computer-aided methods and systems to detect emerging research trends are in increasing demand. Previous efforts have focused on bibliometric analysis. They devoted to the detection of what is emerging, citation counts, amounts of keywords or controlled terms, and the number of authors or publications are used as indicators of emerging research topics (Xu et al., 2021a). However, these citation indicators are not equal to research quality and the accumulation of publications will cause time delay of trends detection (Parraguez et al., 2020). Network analysis is also widely used in trend detection. Co-citation network and bibliograhic coupling are the main stream methods (Porter et al., 2019). Machine learning approaches are also becoming popular (Xu et al., 2019). Another branch of literature applies outliner detection and time series prediction, such as Non-negative Matrix Factorization (NMF) topic modeling method and time series analysis models (Klavans et al., 2020; Weismayer & Pezenka, 2017).

However, the citation and publication delay make the prediction is less timely. To solve the time delay problem, other scientific research-related data sources were introduced. Patent analysis and proposed technical keyword-based analysis of patents were used to monitor emerging technologies (Lee et al., 2018). Patent-paper citation network was used to represent importance indices (Wustmans et al., 2021). Another two biomedical informatics research applied the clinical guideline citation as a gold standard of recognition (Bian et al., 2017, 2019). Inspired by these novel data, we treated the cover papers of journals as important indices for research trend recognition.

Cover image and cover story of an academic journal, to some extent, represent the most important research of a journal. Although the researchers prefer to read online, one article chosen to be prominently placed on the journal's website drew more attention (Wang et al., 2015). The reason why research selected by editors as cover papers is complex: originality, importance, interdisciplinary interest, timeliness, accessibility, elegance, surprising conclusions, or any other factors (Costa & Salvidio, 2020). Some journals, such as Nature Chemistry, claimed that the choice to feature a particular article on the cover of the journal did not imply that the editors think it was better than the other papers in the issue ("Cover story", 2010). However, recently published empirical study demonstrated that both citations and altmetric scores of cover papers were significantly higher than those of non-cover papers (Kong & Wang, 2020). Novel, important and high value research will more likely to lead the research trends (Xu et al., 2021a). For this reason, cover papers of top field journals can be treated as a relative gold standard of research with potential of becoming frontiers.

The widely accepted definition of emerging trends are as follows: (a) radical novelty, (b) relatively fast growth, (c) coherence, (d) prominent impact, and (e) uncertainty and ambiguity (Rotolo et al., 2015). Some researchers found that the greater the extent to which scientific knowledge (a paper) contains emerging ideas, the bigger its scientific impact is (Kwon et al., 2019). These factors are also considered in the decision process of editors and reviewers of the top-ranked journals to publish articles on a particular topic (Uzzi et al., 2013). It is suggested that there are many overlaps between the emerging researches and cover papers published on the top journals. For this reason, we translated the "emerging topics detection" mission to a "cover paper selection" problem. The top journal editors are thought as reliable "gatekeepers" with good judgement of scientific research (Primack et al., 2019). We applied the machine learning technologies to

imitate human intelligence of the top journal (chief) editors, to select the novel, important, and potentially high impact research, and these research have more possibility to become the leader of the research trends. In other words, cover papers of the top journals is a reliable source for emerging topic detection.

Most current research trends detection methods are retrospective, such as citation-based analysis or publication trends analysis. However, even "bursty research topics" (Katsurai & Ono, 2019) detection has time delay, for the reason that emerging research topics formation requires a certain number of publication counts (Fang & Costas, 2020), not to mention citation counts. To overcome this problem, our applied the "real-time" prediction method. Newly-published papers can quickly evaluate whether they have the potential to be selected a cover paper, meaning that they will lead the research trends. Our work is prospective, and it will facilitate the process of emerging trends detection.

Our research focused on the prediction of the potential of being selected as cover paper by top journal (chief) editors. It will, to some extent, to overcome the time delay of the accumulation of publications and citations. And what's more important, our model have the ability to identify the emerging topics potential by learning the research evaluation of top journal editors.

## Related work

### Text mining

Text mining technology is becoming popular in tackling the literature overwhelming problems (Antons et al., 2020). Compared with the traditional bibliometrics-based method, text mining digs into the paper content (Chen, 2005). Topic model algorithm was widely used, and has been applied to detect the emerging research trends (Kawamae & Higashinaka, 2010). Some advanced methods were used to improve the Latent Dirichlet Allocation (LDA) model's performance, such as the Generalized Dirichlet multinomial regression (g-DMR) topic model and structural topic model (STM) (Bai et al., 2020).

By using combination of document influence model (DIM) and citation influence model (CIM), the growth, coherence, and influence indicators were extracted from papers and then were used to train the machine learning (ML) model to predict emerging topics (Xu et al., 2021b). In their following work, DIM was replaced by the topical n-grams (TNG) model to exploit the potential topic models (Xu et al., 2021a). Combined with the topic model, author and citation, author-link topic model, and dynamic author citation topic model were proposed to improve prediction accuracy (Anderson et al., 2012). An improved research grouped the documents by time in each topic to assess the topics' popularity (Li et al., 2018).

Recent years have witnessed the exciting technology "word embedding" in NLP field (Katsurai, 2020). It is often used to identify emergent word meanings, usage, and semantical meaning (Di Carlo et al., 2019; Hamilton et al., 2016; Mihalcea & Nastase, 2012). Recently published Leap2trend (Dridi et al., 2019), applied temporal embeddings and tracked the dynamics of keywords over time, to detect trending scientific topics instantly. And it proved that semantic characterization of research topics yielded better results than keywords for tracing evolving research trends.

### Machine learning

Using the keyword vectors, the k-nearest neighbors (KNN) algorithm was used to predict the forward citations of the patents (Woo et al., 2019). Traditional unsupervised ML technologies, such as PCA, also have good performance in clustering frequently-occurring-together patent records in emerging topics (Wang et al., 2019). Xu et al. (2019) applied the hybrid approaches that combine citation-link and lexical techniques to predict emerging research trends by using several ML models. They also used SVM and random forest to predict important citations (An et al., 2021). Feed-forward multilayer neural network was applied to identify the emerging technologies at early stages by using 18 real-time patent indicators (Lee et al., 2018).

Advanced neural deep learning models became popular for detecting emerging trends providing better performance than existing ML technologies. After building a semantic network on quantum physics, a convolutional neural network (CNN) using states of the semantic network of the past was used to predict future developments (Krenn & Zeilinger, 2020). Using a multimodal approach (including the patent abstract, claims, and indices), researchers built a framework combining of CNN and bidirectional long short-term memory (BiLSTM) in the early detection of valuable patents (Chung & Sohn, 2020). Considering the use of a full citation network, Graph Neural Network (GNN) was used to predict the citation counts, and then identify a set of top papers for predicting research trends (Cummings & Nassar, 2020). Using a manually curated corpus collected from arXiv.org, reinforcement learning and GAN were applied to detect short- and mid-term research trends (Eger et al., 2019). It should be pointed out that the prediction performance of machine learning model has not been well defined.

Link prediction is another widely-used approach in machine learning, which applies the network topology as feature. Network topological features, nodes similarity, and other measures can be utilized to predict the missing links of citation network (Foulds et al., 2015; Takeda & Kajikawa, 2010). Treating the snapshot of keywords network over time as dynamic network, the node centrality measures and node community information were induced to predict the future citation link (Behrouzi et al., 2020). The network topological indicators and dynamics information were used to construct a classification dataset. ML methods were then applied to classify the links and predict future associations among keywords. Autoregressive integrated moving average model (ARIMA), a time series analysis model, was used to predict the values of topological evolution (Choudhury & Uddin, 2016). Main path analysis (MPA) of citation network is another very promising method for understanding the evolution of a scientific domain (Jiang et al., 2020; Xu et al., 2020).

One of the key steps in the process of text classifier model construction is the gold-standard selection. In the process of semi-automated screening of biomedical citations for systematic reviews, research included in the systematic reviews were thought as gold standard of discriminate "relevant" from "irrelevant" citations (Wallace et al., 2010). In the task of relevant references identification, papers cited by the clinical guidelines were thought as the gold standard of research relevance evaluation (Bui et al., 2015). It means the citation of clinical guideline is the most relevant high-quality research in the specific research field (van Dinter et al., 2021).

## Data and methods

### Dataset

Essential Science Indicator (ESI), as an analytical tool in the Web of Science (WoS), is essential for researchers to identify the journals in a certain research area. To better predict the research trends and ignore the discipline factors that may cause confusion to the predicted result, we analyzed papers published in the top-tier journals of the material science discipline. The reason why we chose the material science as our data sample is that it has seen remarkable growth in material science recent years, which not only allows us to collect a larger dataset but also makes it possible for a more precise detection of the research trends. By using the ESI, 23 journals with high quality[1] were selected, 13 journals with few papers and non-cover papers were excluded from the next analytical step.

The remaining 10 journals, including Nature Materials, Nature Nanotechnology, Advanced Materials, Advanced Energy Materials, ACS Energy Letters, Advanced Functional Materials, ACS Nano, Small, Journal of Materials Chemistry A, and Nano Letters were used for in-depth analysis. We retrieved bibliometric information of papers published in these 10 journals from the WoS database during the past ten years (2011–2020), manually encoding a dummy variable "paper type" as 1 if it is a cover paper. The cover papers were collected by browsing the journals' website and identifying the research from every issue. In final, 84,447 papers were obtained, of which 2502 were cover papers and 81,945 were non-cover papers. The cover papers included in this study are collected manually by browsing the journals' websites and reading every issue's cover. All the included cover papers were cross checked by the two authors independently.

### Feature for research trends prediction

To formally predict the research trends, a set of features that are related to the research trend of the paper are included. As is shown in Table 1, these features were divided into three categories, two of which were paper-related and author-related features. The paper-related features include: (1) novelty of research was calculated by using previous approach based on co-citation network of WoS data (Uzzi et al., 2013; Wang et al., 2017), that primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of atypical combinations (Bornmann et al., 2019); (2) international collaborations, which are defined in this study as occurring when authors of paper are from two or more nations; (3) if the paper is highly cited (which is tagged in WoS as "highly cited"); (4) if the paper is a hot paper (which is tagged in WoS as "hot paper"); (5) number of references cited by the paper; (6) title length of the paper; (7) number of keywords of the paper; (8) number of words in the abstract of the paper, and (9) first year citations. The author-related features include: (1) number of authors of the paper; (2) number of addresses of the authors; (3) if the authors are funded; (4) if the fund is a national one,

[1] The included journals are: Nature Reviews Materials, Nature, Science, Nature Materials, Progress in Materials Science, Nature Nanotechnology, Advanced Materials, Materials Science & Engineering R-reports, Materials Today, Advanced Energy Materials, ACS Energy Letters, Advanced Functional Materials, Nano Energy, ACS Nano, International Materials Reviews, Science Advances, Materials Horizons, Nano-Micro Letters, Nature Communications, Small, Journal of Materials Chemistry A, Nano Letters, and Biomaterials.

**Table 1** Definition of features. *Sources* 84,447 papers are retrieved from Web of Science database

| Feature | Definition |
|---|---|
| Paper-related features | |
| Highly cited paper | Dummy, 1 if is the paper is highly cited paper in the WoS database |
| Hot paper | Dummy, 1 if is the paper is hot paper in the WoS database |
| References | Number of references of the paper |
| Pages | Number of pages of the paper |
| Title length | Number of words in the title of the paper |
| Keywords | Number of keywords of the paper |
| Abstract | Number of words in the abstract of the paper |
| Novelty | Value of the novelty score |
| First year citations | Number of citations in the first year after publication |
| Author-related features | |
| Authors | Number of authors of the paper |
| Address | Number of addresses of authors |
| Fund | Dummy, 1 if is the authors receive funds |
| National fund | Dummy, 1 if is the funds the authors receive is national |
| Productivity | Number of papers published by the correspondence author in recent 5 years |
| International collaborations | Dummy, 1 if authors of paper are from two or more nations |
| Altmetric features | |
| Posts | Count mentioned in posts and blogs |
| Tweet | Count mentioned by tweeters |
| MSM | Count mentioned by MSM |
| Accounts | Count mentioned by accounts |
| Feeds | Count mentioned by feeds |
| Patent | Count mentioned by patent |
| Policy | Count mentioned by government documents |

and (5) productivity. The data of these two types of feature categories can be obtained from the WoS database.

Besides the bibliometric information of papers that can be applied to trace the research trends, some online forms such as social media, research blogs, news articles, and policy documents are useful for research trends. These online discussions reflect the research frontier more timely than the citation count, thus, the research trend can be more accurately predicted by taking into account these features. Considering some recently published articles, especially those published in the year 2020, have a lower citation rate in spite of their significance, we calculated how many times these articles were cited by different online sources.

Altmetrics is a better way to understand all potential impacts of scientific research from the online forms (Costas et al., 2015; Priem et al., 2010). The count that papers mentioned by online forms, such as posts, Tweet, MSM, accounts, feeds, patent, and policy documents, can be gathered by altmetrics methods. Altmetrics indicators reduce the time window for predicting the number of citations that need to be accumulated, and the usage of the online altmetrics indicators is easy-going, which can be downloaded and obtained by a simple calculation. Restricted by the data access, we can only collect the summarized altmetrics data from altmetrics.org. To overcome the time accumulation effect, we normalized

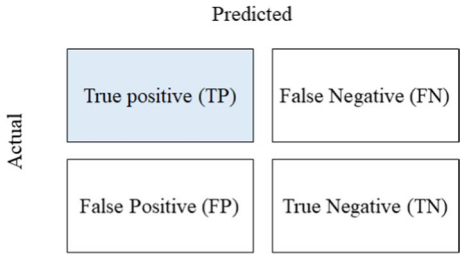


**Fig. 1** Error matrix for machine learning

the altmetrics data by the age of the publication. Because the accumulate velocity of altmetrics indicators are heterogeneous, we normalized the slow source indicators (Feeds, patents and policy) using the mean value calculated by the publication years. And the fast source indicators (Posts, tweets, MSN and accounts) were included directly in the model (Fang & Costas, 2020). Specifically, the R package "rAltmetrics" was used to collect the data (Karthik, 2017). We categorized these 7 features as "altmetric feature" (Table 1).

The affiliations of author, to some extent, represent the team size, which has increasing almost in all areas. Works from larger teams build on popular developments and get the attention immediately (Wu et al., 2019). There is robust correlation that exists between citation impact and team size (Wuchty et al., 2007). The number of keywords, pages and references not only reflect the research productions but also reflect the authors' energies for the publication. It is also found that more references and longer titles will receive more attention (Tan et al., 2016).

## Method for prediction

ML technologies with access to a large dataset could discover and predict research trends using feasible features. We use the supervised ML methods to effectively detect the research trends. To be specific, we used seven ML technologies (Gaussian naive Bayes, AdaBoost, random forest, logistic regression, k-nearest neighbors, Calibrated ClassifierCV, and decision tree), which are more widely used method developed specifically for the prediction task. The prediction is built from the scientific papers we collected, using two-thirds of the sample as a training data set to find the best parameters for each of the ML algorithm. Then, the well-studied algorithm was applied to the 33% test data set to evaluate the rate of prediction.

To comprehensively evaluate the performance of predicting the research trend of these ML methods, miss errors in which the cover papers were not detected are adopted. Typically, we used recall score, the gold standard of approach for evaluating miss errors, to detect the proportion of cover papers.

The recall score is a criterion for evaluating true positive rate, a higher recall rate means more cover papers can be correctly predicted. As the error matrix of ML in Fig. 1, the recall score rate and precision rate are calculated by the formulas:

$$\text{Recall rate} = \frac{TP}{TP + FN},$$

$$\text{Precision rate} = \frac{TP}{TP + FP}$$

The aim of this study is to evaluate the seven selected ML models, finding the best model with a higher recall rate which is beneficial to provide a user-friendly approach for the detection of research trends.

To compare the accuracy performance, the area under the receiver-operating-characteristic (ROC) curve (AUC) was used (Ling et al., 2003). The ROC curve was also adopted to compare the ML models' performance across the entire range of error rates in each of the models (Bradley, 1997).

To ensure the best baseline performance of the ML algorithm, the grid search method was used to find the optimal parameters. Tenfold cross-validation was conducted to training the data, which provide the optimal parameter combination. By implementing the "GridSearchCV" in the sklearn package (Pedregosa et al., 2011), we systematically searched the optimal parameter of each ML technology. To be consistent with the purpose of finding the best ML model with a higher recall score, the optimal parameter was determined by the best recall score on the test dataset. The optimal parameters of the seven ML models are presented in Table 2.

To evaluate the generalization ability of our well-trained models, we further collected data from three most high-impact journals published in 2019, including Cell, Nature, and Science. All the bibliometrics data were also collected and cover papers were manually identified. The CNS sample contained 1622 papers, 118 of which was cover paper. We tested the recall score of the sample to check the externality of the trained ML models, checking if the models in terms of detecting research trends performed well in these journals as well.

## Results

### Descriptive analysis

Table 3 presents the descriptive statistical analysis of 17 variables, and five dummy variables were excluded in this analysis.

### Topic distribution result

We formed the categories of topics of the papers we collected by applying computer-assist techniques–topic model to check the topic distribution of the cover papers. Specifically, a probabilistic topic model called LDA was employed to automatically identify main topics (the optimal number of topics is 19) from each abstract of the papers (Blei et al., 2003). The underlying assumption of LDA is that the abstract of a paper consists of multiple topics, each topic is a discrete probability distribution over words (Steyvers & Griffiths, 2007). Thus, the topic name (which are manually labeled) could be inferred based on algorithmically found top words.

Before the text mining analysis, preliminary data processing process (including words stemmization, words lowercase, punctuation, numbers and stop words removing) were conducted to ensure the clusters of topics will have significant meanings. To determine the number of topics in our analysis, four metrics in R package "ldatuning" were also used to examine the model performance and to estimate a reasonable number of topics (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths & Steyvers, 2004). We found that

**Table 2** Optimal parameters for the seven machine learning models

| Models_Parameters | Parameters description | Optimal value |
|---|---|---|
| GaussianNB | | |
| var_smoothing | Parameter that determine the calculation stability | 4 |
| AdaBoost | | |
| learning_rate | Parameters that determine the learning speed | 1 |
| n_estimators | The number of estimators | 200 |
| Random forest | | |
| criterion | The function to measure the split quality | Entropy |
| max_depth | The number of maximum depths of the tree | 10 |
| max_features | The number of features that search for the best split | 4 |
| min_samples_split | The number of minimum samples to split an internal node | 4 |
| min_samples_leaf | The number of minimum samples required to be a leaf | 5 |
| n_estimators | The number of estimators | 200 |
| Logisitic regression | | |
| penalty | The penalization function | l2 |
| C | The inverse of regularization strength | 1 |
| KNearest | | |
| n_neighbors | The number of neighbors | 35 |
| weights | Weight function used in prediction | Distance |
| metric | The distance metric to use for the tree | Euclidean |
| leaf_size | Parameter that affect the construction, query, and the memory to store the tree | 5 |
| P | Power parameter for the Minkowski metric | 2 |
| CalibratedClassifierCV | | |
| penalty | The penalization function | l2 |
| C | The inverse of regularization strength | 1 |
| max_iter | The maximum iteration for the algorithm convergence | 3000 |
| Decision tree | | |
| criterion | The function to measure the split quality | Entropy |
| max_depth | The number of maximum depths of the tree | 13 |
| min_samples_split | The number of minimum samples to split an internal node | 3 |
| min_samples_leaf | The number of minimum samples required to be a leaf | 7 |

19 topics best average performance across the four metrics. Thus, we clustered the 84,447 abstracts of papers into the optimal 19 topics.

We then classified each paper into the topic with the highest probability distribution. The descriptive statistics of the cover paper were presented in Table 4. It can be seen that the proportion of cover paper in each topic is around 3%, which is consistent with the proportion of those in the sample (2.96%). The result suggests that the topic distribution of the cover papers is equal, cover papers normally distribute in the research topic in the material science discipline.

We also conducted a sub-group prediction analysis by using the 19 topic models to produce clusters separately. We adapted the model including overall the features, and found that the performance was consistent in all the sub-clusters. The average recall rate was 0.68, precision rate was 0.56, and the AUC is 0.71, some prediction results (4/19) even

**Table 3** Description of features and outcome variable. *Sources* Data of paper-related feature and author-related feature are retrieved from Web of Science database. Altmetric features are collected using R package "rAltmetrics"

| Feature | N | Mean | SD | Median | Min | Max |
|---|---|---|---|---|---|---|
| Paper-related features | | | | | | |
| References | 84,447 | 47.63 | 19.16 | 76 | 0 | 793 |
| Pages | 84,447 | 7.89 | 4.12 | 8 | 1 | 71 |
| Title length | 84,447 | 11.12 | 3.11 | 10 | 2 | 37 |
| Keywords | 84,447 | 6.95 | 2.19 | 6 | 1 | 10 |
| Abstract | 84,447 | 143.57 | 32.99 | 187 | 1 | 557 |
| Citation in the first year | 84,447 | 14.44 | 7.16 | 6 | 0 | 279 |
| Novelty | 84,447 | 2.16 | 1.70 | 0.93 | 0 | 6.988 |
| Author-related features | | | | | | |
| Authors | 84,447 | 8.15 | 2.44 | 5 | 1 | 86 |
| Address | 84,447 | 4.59 | 2.59 | 3 | 1 | 86 |
| Productivity | 84,447 | 4.63 | 2.45 | 6 | 0 | 17 |
| Altmetric features | | | | | | |
| Posts | 39,667 | 9.94 | 59.27 | 2 | 1 | 9286 |
| Tweet | 33,832 | 5.56 | 43.71 | 2 | 1 | 6328 |
| MSM | 6439 | 7.39 | 11.37 | 3 | 1 | 308 |
| Accounts | 31,154 | 4.43 | 53.09 | 2 | 1 | 6701 |
| Feeds | 2976 | 1.29 | 1.06 | 1 | 1 | 45 |
| Patent | 3189 | 2.98 | 1.98 | 1 | 1 | 70 |
| Policy | 31,146 | 0.01 | 0.04 | 0 | 0 | 1 |

exceeded the best performance of the prediction model. This could be explained partly by the topic specific characters of the prediction model.

## Evaluation result of machine learning models

Considering the data is unbalanced, the proportion of cover paper is small, it will introduce bias during the training process. As a result, the training rest will bring about the true positive issue, most of the cover paper will not be predicted correctly. To solve this issue and minimize the impact of the sample bias when we randomly split the data into training and test data sets, an approach for constructing the classifiers from unbalanced data-synthetic Minority Oversampling Technique (SMOTE) is used (Chawla et al., 2002). SMOTE creates a "synthetic" minority class by using the over-sampling approach, it generates synthetic cover paper samples so that a balanced training data set is created. We used the papers published in 2016 as training sets, in which the number of cover papers is only 248, while the number of non-cover papers is 7,074. When using SMOTE, the number of cover papers in the training data set is equal to that of non-cover paper, with the number of 7074. The other publications published in 2011–2015 and 2017–2020 were used to evaluate the prediction performance.

Figure 2 displays the ROC curve of the seven ML models. We first excluded the "novelty", "Altmetrics" and "citation in the first year" from all the features, and then added them one by one (Fig. 2a–c). It could be found that the Altmetric features indeed

**Table 4** Summary statistics of 19 topics

| Topic | Cover paper | | Number of topic (n) |
|---|---|---|---|
| | Number (n) | Proportion in topic (%) | |
| Topic_1 | 97 | 4.31 | 2251 |
| Topic_2 | 205 | 2.64 | 7765 |
| Topic_3 | 118 | 3.76 | 3138 |
| Topic_4 | 113 | 2.42 | 4669 |
| Topic_5 | 127 | 2.47 | 5141 |
| Topic_6 | 194 | 4.56 | 4254 |
| Topic_7 | 88 | 2.07 | 4251 |
| Topic_8 | 134 | 3.84 | 3490 |
| Topic_9 | 156 | 3.48 | 4483 |
| Topic_10 | 117 | 2.25 | 5200 |
| Topic_11 | 94 | 4.29 | 2191 |
| Topic_12 | 258 | 2.15 | 12,002 |
| Topic_13 | 129 | 3.32 | 3886 |
| Topic_14 | 153 | 3.61 | 4238 |
| Topic_15 | 96 | 2.09 | 4593 |
| Topic_16 | 142 | 2.88 | 4931 |
| Topic_17 | 120 | 3.92 | 3061 |
| Topic_18 | 79 | 2.79 | 2832 |
| Topic_19 | 82 | 3.96 | 2071 |
| All papers | 2,502 | 2.96 | 84,447 |

The probabilistic topic model LDA was used to cluster the abstracts of 84,447 papers into 19 topics. The LDA results of 2502 cover papers were presented. *Number (n)* denotes the number of cover papers in each LDA topic; *Number of topic (n)* denotes the number of papers in each LDA topic; *Proportion in topic (%)* is the proportion of cover papers in each topic.

slightly improved the performance of prediction models, while the citation feature lessened the performance. It is hinted that citation based indicators could not make sense. Additionally, it is worth noting that the Logistic regression and Linear SVM model with the best performance in the most tests. The AUC of the prediction model including overall features is 0.7, as is shown in Table 5 with the recall rate and the precision rate. We also tested the prediction performance by using datasets 2011–2015 and 2017–2020, separately. We found that the AUCs of 2017–2020 are higher than datasets of 2011–2015 (Appendix Fig. 4).

In this research, we preferred the high performance of the recall rate. The false negative rate (non-cover paper predicted as cover paper) was considered with lower priority. It means that we would include some "false positive" papers, which is the trade-off between the high recall rate and precision rate.

When excluding international collaborations and productive features, we found that the performance decreased (Recall = 0.63, Precision = 0.37, AUC = 0.61), which means the two author-related features are important. Although a research investigated Scopus data of 2005 found that international collaboration tends to produce conventional research (Wagner et al., 2019), a recently study focused on COVID-19 field reached the opposite

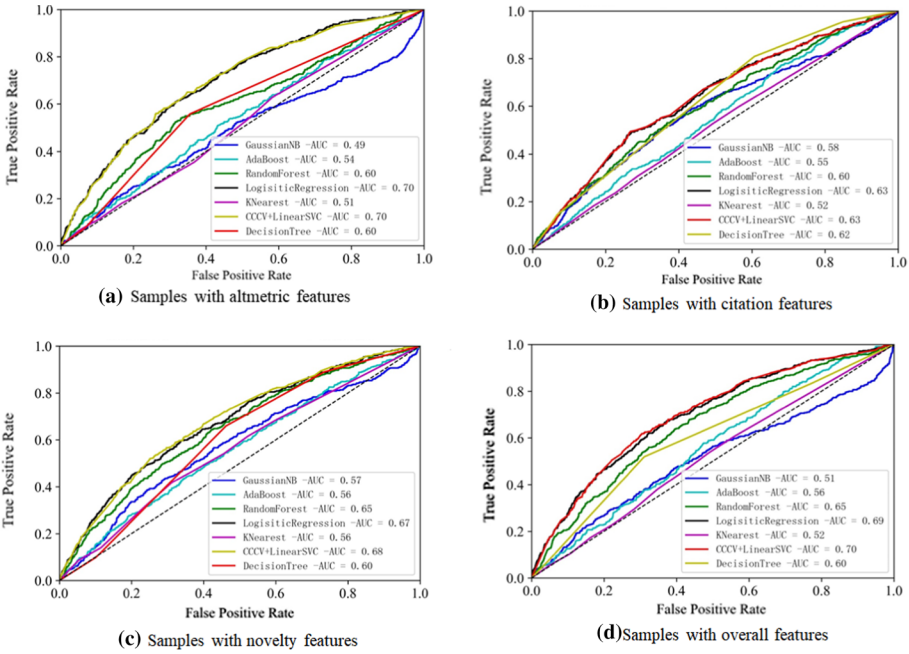


**Fig. 2** ROC of the seven machine learning models

**Table 5** Prediction performance of 7 machine learning models with overall features

| Models | Recall | AUC | Precision |
|---|---|---|---|
| GaussianNB | 0.61 | 0.51 | 0.32 |
| AdaBoost | 0.67 | 0.56 | 0.37 |
| Random forest | 0.71 | 0.65 | 0.42 |
| Logistic regression | 0.76 | 0.69 | 0.47 |
| KNearest | 0.64 | 0.52 | 0.34 |
| CalibratedClassifierCV | 0.77 | 0.70 | 0.48 |
| Decision tree | 0.69 | 0.60 | 0.39 |

The prediction performance results with almetric features, citation features, and novelty features are presented in Appendix Table 6.

conclusion (Liu et al., 2021). Our study of materials science supported, to some extent, the later opinion, for the reason that this filed is labor intensive and device induced.

## Externality of the machine learning models

The ROC curve of the CNS papers is present in Fig. 3. It can be seen that the performance of the logistic regression (recall = 0.71, precision = 0.51) and the linearSVC (recall = 0.73, precision = 0.49) are better. The generalization of the prediction model was consistent with that of the samples in the material science discipline. It indicates that our ML method in terms of identification of cover paper could be applied in a wide range of samples.

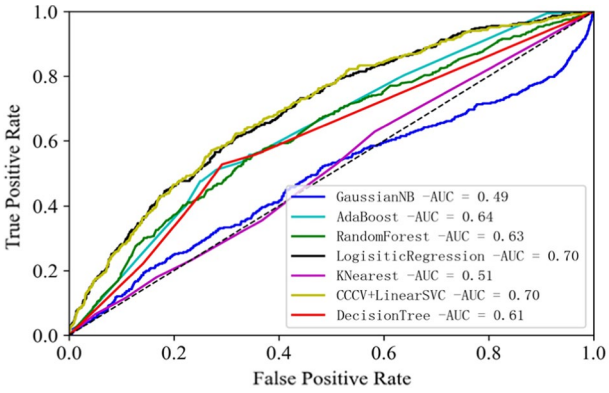


**Fig. 3** The ROC curve of the CNS papers

Moreover, the prediction performance of external data was better than that in material science, which could be explained by the three journals are all high-quality.

## Validation of field expert review

The prediction of results was validated by 10 experts from materials science. We mapped the prediction results to previous topic model clustered groups. Experts reviewed the results and concluded that the papers we presented included almost all the emerging topics in recent years. It is also highlight that our prediction results in 2019 covered all the emerging topics, which means the framework we proposed overcome the time delay problem of traditional emerging topics detection.

Ten filed experts were invited to evaluate the model performance by choosing the cover papers in two pre-designed tests. The experts are currently working in the material science department and physical science department of two universities. Their research interests are mental materials, material chemistry, electronic materials, optical material, computational materials, nanomaterial and nanotechnology, and theoretical physics. All the experts (3 professors, 3 associated professors and 4 Ph.D. Candidates) published more than at least ten papers in the materials field. Before the formal test, every expert selected 3 topics, and was assigned 20 cover papers and 80 non-cover papers to train their ability of cover paper identification. It was ensured that papers belong to every topic has two "reviewers" to evaluate whether it could be selected as cover paper. After that, we conducted a test accuracy experiment to compare the performance of the prediction model and the experts. Cohen's κ statistic was used to evaluate the experts' inter-rater reliability.[2]

Ten experts were assigned to identify the 10 cover papers from 100 papers published in 10 materials journals in 2021. Innovation and importance were considered firstly in the process of evaluation. The cover papers and non-cover papers are all from the topics they selected. The tests were conducted 4 times and every expert evaluated more than 400 papers, which ensured that every paper was reviewed by two experts independently.

[2] κ = (0–0.20) means poor agreement, (0.21–0.40) means slight agreement, (0.41–0.60) means fair agreement, (0.61–0.80) means good agreement, (0.81–0.92) means very good agreement, and (0.93–1) means excellent agreement.

The average recall rate of the prediction model was 0.734 and the precision rate was 0.690, which was higher than the average recall rate of experts at 0.595 and 0.673, respectively. The agreement of the two experts was slight consistency ($\kappa = 0.58$). Considering the experts have priori knowledge, such as could have read some papers before, prediction model would still perform on par with the experts. In addition to a little higher performance, the prediction model has also been consistency of producing results.

## Discussion and conclusions

This study provides a compressive ML framework to predict the emerging research trends in an innovative way. It is useful to overcome the time delay problem of traditional publication counts based methods. We have developed a principled and extensible approach for identifying leader papers of emerging topics by combining the topic model and prediction models. By consulting with 10 field experts, the cover papers we selected are proved that they included all the emerging topics. This result ensured the effectiveness of our approach to detect emerging papers instantly. What's more, our model applied some easily collected bibliometrics data and this will helpful to simplify a great deal of the pre-processing works. The comparison test has shown that prediction model has comparable performance with materials science experts. Additionally, the non-fatigue characteristic of prediction model enables constant training and learning until achieving satisfactory accuracy. It could reduce the cognitive burden on researchers, and will helpful to increase the efficiency of understanding the research domains.

The strategy presented here is based on the hypothesis that paper featured on the cover of a top journal has more possibility to become a leader paper of an emerging topic. Cover papers are closely related to leader papers of emerging topics in terms of relatively fast growth of publications and citations, radical novelty and prominent impact. The process of cover paper selection is, to some extent, a kind of "knowledge distillation" in essence (Gou et al., 2020). The ML models we proposed can be thought as a top editor, whose has a good scientific taste and judgement of identifying important and novel papers. It is also critical to understand that the cover paper of top journal is not quite equal to the leader paper of emerging topics. However, when facing with the challenge of literature flood, our approach can reduce the time of citation accumulation. And this might seem to be an acceptable solution of information consolidation.

Although the idea of the prediction model we constructed stemmed from the cover paper selection, the readers of a newly-published paper always passively get some related information, such as citation counts or discussions on the twitter in the real world. The model, to some extent, is a post-publication peer review, rather than editors' review. Additionally, it is suggested that over half of altmetric events happened within 1 month after the research published (Fang & Costas, 2020).This makes the post-publication "review" become more efficient. It is noted that, in this research, we preferred the high performance of the recall rate. The false negative rate (non-cover paper predicted as cover paper) was considered with lower priority. It means that we would include some "false positive" papers, which is the trade-off between the high recall rate and precision rate.

Our analysis also showed some important results. First, Altmetrics has effectively replace the citation counts in the prediction models. The precision performance of Altmetrics based model was significantly higher than citation counts based model. Considering one aim of this study is focused on reducing the time delay of data collection, Altmetrics

is a good indicator, especially in those research trends prediction mission. Thus, our results suggest that feature engineering of bibliometrics data would probably be an impractical endeavor. It is suggested that over half of altmetric events happened within 1 month after the research published (Fang & Costas, 2020). This makes the post-publication "review" become more efficient. More deep learning model should be introduced to avoid feature selection.

This study inevitably has some limitations. As pointed out previously, our approach is incomplete despite the promise inherent in the circumstantial prediction. The expert reviewed our results found that our model predicts many unrelated papers. This is owing that we pursue the highest precision rate of our model. Secondly, although the approach has been applied in biological journals and top interdisciplinary journals, more work should be conducted to ensure the fitness of our prediction models. Some cases, such as sleeping beauty literature should be used to validate our model to evaluate the ability of weak signal detection (Ke et al., 2015).

Future research will focus on the identifying the causal relationship between cover letter and emerging research trends. Cover papers have some advantages of high visibility, and this will make readers have more interests to conduct follow-up research. Is there a bidirectional casualty between them? And how the evaluation dynamics of the relationship between them? Indeed, our model should be extended to consider the dynamics of topics, and this will useful in sub-topic level emerging trends detection. Another interesting challenge for a long-term research topic would be the comparison of our methods with traditional prediction models in handling weak signals of emerging topics. To ensure the fairness of assessment, the performance evaluation framework is ugly needed.

## Appendix

See Table 6 and Fig. 4.

**Table 6** Prediction performance of 7 machine learning models with altmetric features, citation features, and novelty features

| Models | Altmetric features | | | Citation features | | | Novelty features | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | AUC | Precision | Recall | AUC | Precision | Recall | AUC | Precision |
| GaussianNB | 0.54 | 0.49 | 0.33 | 0.61 | 0.58 | 0.35 | 0.65 | 0.57 | 0.34 |
| AdaBoost | 0.65 | 0.54 | 0.33 | 0.59 | 0.55 | 0.33 | 0.65 | 0.56 | 0.33 |
| RandomForest | 0.68 | 0.60 | 0.40 | 0.64 | 0.60 | 0.40 | 0.72 | 0.65 | 0.39 |
| LogisiticRegression | 0.76 | 0.70 | 0.47 | 0.69 | 0.63 | 0.40 | 0.76 | 0.67 | 0.44 |
| KNearest | 0.62 | 0.51 | 0.34 | 0.51 | 0.52 | 0.34 | 0.62 | 0.56 | 0.31 |
| CCCV + LinearSVC | 0.76 | 0.70 | 0.47 | 0.70 | 0.63 | 0.41 | 0.77 | 0.68 | 0.46 |
| DecisionTree | 0.67 | 0.60 | 0.39 | 0.66 | 0.62 | 0.39 | 0.70 | 0.60 | 0.35 |

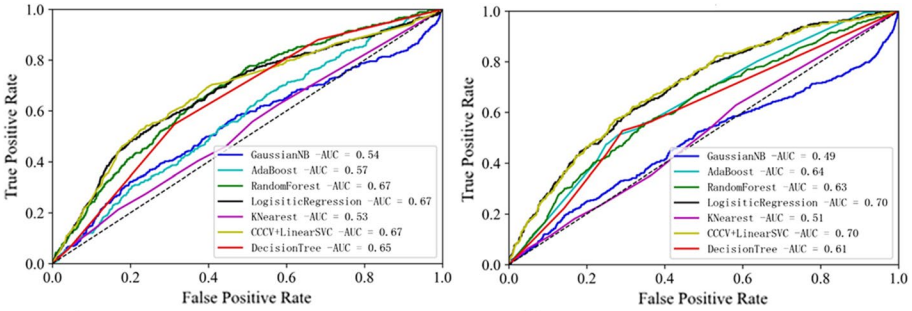


**Fig. 4** ROC of the seven machine learning models using datasetes 2011–2015, 2017–2020

**Acknowledgements** This work was supported by the China Scholarship Council.

**Author contributions** WW: Processed the experimental data, performed the analysis, Contribution to the overall paper, Other contribution. HL: Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the manuscript. ZS: Conceived and designed the analysis, Wrote the manuscript and designed the figures, Other contribution.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

An, X., Sun, X., Xu, S., Hao, L., & Li, J. (2021). Important citations identification by exploiting generative model into discriminative model. *Journal of Information Science, 016*, 5551.

Anderson, A., Jurafsky, D., & McFarland, D. (2012). Towards a computational history of the acl: 1980–2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (pp. 13–21).

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management, 50*(3), 329–351.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402).

Bai, X., Zhang, X., Li, K. X., Zhou, Y., & Yuen, K. F. (2020). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*.

Behrouzi, S., Sarmoor, Z. S., Hajsadeghi, K., & Kavousi, K. (2020). Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics, 14*(4), 101079.

Bian, J., Abdelrahman, S., Shi, J., & Del Fiol, G. (2019). Automatic identification of recent high impact clinical articles in PubMed to support clinical decision making using time-agnostic features. *Journal of Biomedical Informatics, 89*, 1–10.

Bian, J., Morid, M. A., Jonnalagadda, S., Luo, G., & Del Fiol, G. (2017). Automatic identification of high impact articles in PubMed to support clinical decision making. *Journal of Biomedical Informatics, 73*, 95–103.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222.

Bornmann, L., Tekles, A., Zhang, H. H., & Fred, Y. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13*(4), 100979.
Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.
Bui, D. D. A., Jonnalagadda, S., & Del Fiol, G. (2015). Automatically finding relevant citations for clinical guideline development. *Journal of Biomedical Informatics, 57*, 436–445.
Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Chen, C. (2005). *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Wiley.
Choudhury, N., & Uddin, S. (2016). Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics, 108*(2), 745–776.
Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change, 158*, 120146.
Costa, A., & Salvidio, S. (2020). Animal behaviour on the cover: Layout cover patterns of ethological journals. *Ethology Ecology & Evolution, 1*, 1–9.
Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology, 66*(10), 2003–2019.
Cover Story. (2010). *Nature Chemistry*, *2*(3), 147. https://doi.org/10.1038/nchem.555
Cummings, D., & Nassar, M. (2020). Structured citation trend prediction using graph neural networks. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3897–3901).
Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61–84.
Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). Training temporal word embeddings with a compass. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 6326–6334).
Dridi, A., Gaber, M. M., Azad, R. M. A., & Bhogal, J. (2019). Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access, 7*, 176414–176428.
Eger, S., Li, C., Netzer, F., & Gurevych, I. (2019). Predicting research trends from arxiv. http://arxiv.org/abs/1903.02831.
Fang, Z., & Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics, 123*(2), 1077–1101.
Foulds, J., Kumar, S. H., & Getoor, L. (2015). Latent topic networks: A versatile probabilistic programming framework for topic models. *Proceedings of the 32nd International Conference on Machine Learning*, *37*(2003), 777–786. http://linqs.cs.umd.edu/basilic/web/Publications/2015/foulds:icml15/
Gibney, E. (2014). How to tame the flood of literature. *Nature News, 513*(7516), 129.
Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2020). *Knowledge distillation: A survey*. http://arxiv.org/abs/2006.05525.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5228–5235.
Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing conference on empirical methods in natural language processing* (Vol. 2016, p. 595).
Jiang, X., Zhu, X., & Chen, J. (2020). Main path analysis on cyclic citation networks. *Journal of the Association for Information Science and Technology, 71*(5), 578–595.
Karthik, R. (2017). rAltmetric: Retrieves altmerics data for any published paper from altmetrics.com.http://CRAN.R-project.org/package=rAltmetric.
Katsurai, M. (2020). Using word embeddings for library and information science research: A short survey. *ACM SIGWEB Newsletter, 1*, 1–7.
Katsurai, M., & Ono, S. (2019). TrendNets: Mapping emerging research trends from dynamic co-word networks via sparse representation. *Scientometrics, 121*(3), 1583–1598.
Kawamae, N., & Higashinaka, R. (2010). Trend detection model. In *Proceedings of the 19th international conference on World wide web* (pp. 1129–1130).
Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences, 112*(24), 7426–7431.
Klavans, R., Boyack, K. W., & Murdick, D. A. (2020). A novel approach to predicting exceptional growth in research. *PLoS ONE, 15*(9), e0239177.

Kong, L., & Wang, D. (2020). Comparison of citations and attention of cover and non-cover papers. *Journal of Informetrics, 14*(4), 101095. https://doi.org/10.1016/j.joi.2020.101095
Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences, 117*(4), 1910–1916.
Kwon, S., Liu, X., Porter, A. L., & Youtie, J. (2019). Research addressing emerging technological ideas has greater scientific impact. *Research Policy, 48*(9), 103834.
Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.
Li, C., Feng, S., Zeng, Q., Ni, W., Zhao, H., & Duan, H. (2018). Mining dynamics of research topics based on the combined LDA and WordNet. *IEEE Access, 7*, 6386–6399.
Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A better measure than accuracy in comparing learning algorithms. In *Conference of the Canadian society for computational studies of intelligence* (pp. 329–341).
Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., Freeman, R. B., Meyer, E. T., Yoon, W., Sung, M., & Jeong, M. (2021). Pandemics are catalysts of scientific novelty: Evidence from COVID-19. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24612
Mihalcea, R. & Nastase, V. (2012). Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 259–263).
Parraguez, P., Škec, S., Carmo, D. O., & Maier, A. (2020). Quantifying technological change as a combinatorial process. *Technological Forecasting and Social Change, 151*, 119803.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.
Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change, 146*, 628–643.
Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Springer.
Primack, R. B., Regan, T. J., Devictor, V., Zipf, L., Godet, L., Loyola, R., Maas, B., Pakeman, R. J., Cumming, G. S., Bates, A. E., & Pejchar, L. (2019). *Are scientific editors reliable gatekeepers of the publication process?* Elsevier.
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827–1843.
Salatino, A. A. (2019). Early Detection of Research Trends. *CoRR*, *abs/1912.0*. http://arxiv.org/abs/1912.08928
Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis, 427*(7), 424–440.
Takeda, Y., & Kajikawa, Y. (2010). Tracking modularity in citation networks. *Scientometrics, 83*(3), 783–792.
Tan, L. S. L., Chan, A. H., & Zheng, T. (2016). Topic-adjusted visibility metric for scientific articles. *Annals of Applied Statistics, 10*(1), 1–31. https://doi.org/10.1214/15-AOAS887
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013a). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.
van Dinter, R., Catal, C., & Tekinerdogan, B. (2021). A decision support system for automating document retrieval and citation screening. *Expert Systems with Applications, 182*, 115261.
Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy, 48*(5), 1260–1270.
Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics, 11*(1), 1–11.
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, X., Liu, C., & Mao, W. (2015). Does a paper being featured on the cover of a journal guarantee more attention and greater impact? *Scientometrics, 102*(2), 1815–1821.
Wang, Z., Porter, A. L., Wang, X., & Carley, S. (2019). An approach to identify emergent topics of technological convergence: A case study for 3D printing. *Technological Forecasting and Social Change, 146*, 723–732.
Weismayer, C., & Pezenka, I. (2017). Identifying emerging research fields: A longitudinal latent semantic keyword analysis. *Scientometrics, 113*(3), 1757–1785.
Woo, H.-G., Yeom, J., & Lee, C. (2019). Screening early stage ideas in technology development processes: A text mining and k-nearest neighbours approach using patent information. *Technology Analysis & Strategic Management, 31*(5), 532–545.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378–382.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.

Wustmans, M., Haubold, T., & Bruens, B. (2021). Bridging trends and patents: Combining different data sources for the evaluation of innovation fields in blockchain technology. *IEEE Transactions on Engineering Management*.

Xu, H., Winnink, J., Yue, Z., Zhang, H., & Pang, H. (2021a). Multidimensional Scientometric indicators for the detection of emerging research topics. *Technological Forecasting and Social Change, 163*, 120490.

Xu, S., Hao, L., An, X., Pang, H., & Li, T. (2020). Review on emerging research topics with key-route main path analysis. *Scientometrics, 122*(1), 607–624.

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*(4), 100983.

Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021b). A topic models based framework for detecting and forecasting emerging technologies. *Technological Forecasting and Social Change, 162*, 120366.