




# A deep learning approach for identifying biomedical breakthrough discoveries using context analysis

Xue Wang<sup>1</sup> · Xuemei Yang<sup>1</sup> · Jian Du<sup>2</sup> · Xuwen Wang<sup>1</sup> · Jiao Li<sup>1</sup> · Xiaoli Tang<sup>1</sup> 

Received: 26 June 2020 / Accepted: 17 April 2021 / Published online: 23 May 2021  
© The Author(s) 2021

## Abstract

Breakthrough research in scientific fields usually comes as a manifestation of major development and advancement. These advances build to an epiphany where new ways of thinking about a problem become possible. Identifying breakthrough research can be useful for cultivating and funding further innovation. This article presents a new method for identifying scientific breakthroughs from research papers based on cue words commonly associated with major advancements. We looked for specific terms signifying scientific breakthroughs in citing sentences to identify breakthrough articles. By setting a threshold for the number of citing sentences (“citances”) with breakthrough cue words that peer scholars often use when evaluating research, we identified articles containing breakthrough research. We call this approach the “others-evaluation” process. We then short-listed candidates from the selected articles based on the authors’ evaluations of their own research, found in the abstracts. This we call the “self-evaluation” process. Combining the two approaches into a dual “others-self” evaluation process, we arrived at a sample of 237 potential breakthrough articles, most of which are recommended by the Faculty Opinions. Based on the breakthrough articles identified, using SVM, TextCNN, and BERT to train the models to identify abstracts with breakthrough evaluations. This automatic identification model can greatly simplify the process of others-self-evaluation process and promote identifying breakthrough research.

**Keywords** Deep learning · Context analysis · Breakthrough discoveries · Breakthrough identification · Citances

## Introduction

Science is a dynamic system of scientific advances in which breakthrough discoveries are important developmental markers whose impact may extend beyond their own field of study. From the perspective of scientific research and policy, breakthrough discoveries

---

✉ Xiaoli Tang  
tang.xiaoli@imicams.ac.cn

<sup>1</sup> Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100005, China

<sup>2</sup> National Institute of Health Data Science, Peking University, Beijing 100191, China

have attracted the interest of researchers in many fields. In fact, several scholars have even established theoretical models to identify and explore the features of breakthrough research (Chen, 2012; Chen et al., 2009; Ponomarev et al., 2014; Winnink et al., 2016; Wolcott et al., 2016). At the same time, to accelerate scientific development, governments and funding agencies around the world have invested considerable resources into prioritizing and funding emerging research—especially transformative research. In 2007, the American National Science Board proposed creating programs to enhance support for transformative research (National Science Board, 2007). More recently, the US National Institutes of Health (2016) established the Transformative Research Projects Program (R01), which specifically targets transformative research that aims to support “exceptionally innovative and/or unconventional research projects with the potential to create or overturn fundamental paradigms”. In 2017, the Report of the 19th National Congress of the Communist Party of China (2017) also promoted these ends, emphasizing the need to make China a country of innovators who could contribute major breakthroughs to pioneering basic research. Although each country has established funding and research infrastructures to foster breakthrough research, determining how to best utilize funding and which projects to fund is still an important challenge. Significant gains in scientific research could result for both scientists and agencies if breakthrough research could be identified early on and promoted through sympathetic funding policies. Therefore, our purpose with this research was to develop a feasible method of filtering breakthrough discoveries out of the vast expanses of academic literature—ideally before they have become “old news”.

## What is a breakthrough?

At present, there is no uniform definition of a breakthrough, which obviously makes it difficult to identify one. Hollingsworth (2008) defines a breakthrough as “a finding or process, often preceded by numerous small advances, which leads to a new way of think about a problem...”. He argues that an essential property of a breakthrough is “...new way of thinking about a problem”. Another significant feature of breakthroughs is their link to creative, transformative and groundbreaking research (Ponomarev et al., 2014; Winnink et al., 2019; Wolcott et al., 2016). According to Kuhn’s theory of scientific development (Kuhn, 1962), periods of “conventional science” can be interrupted by “scientific revolutions” which transform science and result in a new stage of “conventional science”. During times of conventional science, incremental innovation supplements previous research under the existing paradigm, promoting the cumulative development of science. But during the relatively rare occurrences of a “scientific revolution”, science evolves rapidly through new insights, research methods or theoretical explanations, overthrowing the old paradigm in favor of a new framework. Rather than operating independently, there is a dynamic and interactive relationship between incremental research and scientific breakthroughs, with the gradual accumulation of incremental innovations ultimately leading to a new scientific paradigm.

The American National Science Board and National Science Foundation—two major scientific decision-making and funding agencies—champion the notion of “transformational research” as the means by which breakthroughs are made in scientific research. While the National Science Board (2007) defines transformative research in terms of policy, and the National Science Foundation (2015) identifies it through the management perspective of research funding, both believe that such research has the potential to lead

to paradigm shifts. In that sense, transformative research has a similar connotation to the notion of a scientific revolution in Kuhn's theory.

Transformative research, therefore, has the essential attributes of breakthrough research. At the same time, while important scientific discoveries produced by incremental research may not have sufficient transformative potential, they can provide new ideas and knowledge that play a crucial role in breakthrough research. As such, we regard both transformative research and major incremental scientific innovations as breakthroughs, and seek methods by which to recognize either and both in scientific publications.

### **Prior work identifying breakthrough publications**

The current approaches to identifying breakthroughs mostly fall into two categories. One is the qualitative identification method used by academic communities, known as peer review. The other is based in scientometrics.

Peer review is a comprehensive evaluation approach that, no matter what automatic method is invented, will remain an important and effective method by which such breakthroughs are identified. But, it also has obvious shortcomings, being both time-consuming and highly dependent on expert opinions. Further, given the explosive growth of scientific literature, peer review is becoming an increasingly inefficient means by which to identify the presence of valuable breakthrough research in publications.

Most scientific research to identify breakthrough publications therefore falls within the field of scientometrics. Methods of identifying breakthroughs often prioritize citation statistics. High citation counts are used to identify high-value articles and even predict Nobel Prize winners (Garfield & Welljams-Dorof, 1992). Recently, however, techniques such as that developed by Ponomarev et al. (2014) are formulating single indicators for the early detection of candidate breakthroughs based on the dynamics of publication citation. Wolcott et al. (2016) incorporated multiple time-dependent and time-independent features of publications into a model to differentiate known breakthrough research from randomly selected control papers, such as the key publications reported in a high-quality data set like The American Society of Clinical Oncology (ASCO) Annual Report. Some researchers have used cited or long-term reference analysis to detect publications containing seminal research (Comins & Hussey, 2015) or research milestones (Comins & Leydesdorff, 2017). The idea is that transformative research appears to cause a “disruption” in the citation chain of the prevailing research paradigm. Hence, papers a given a “disruption score” and those exceeding a threshold may contain breakthrough research. The technique has been tested successfully with research in the fields of physics, computer science, and biomedicine (Huang et al., 2013, 2014). Winnink et al. (2016, 2019) demonstrate that characteristic patterns in the citation profiles of known breakthrough publications can be used in the early detection of discoveries with an important impact on scientific development. However, identifying breakthrough research using citation statistics relies on a correlation between breakthrough publications and external indicators rather than causation. For example, not all citations are equally important or even positive (Hernandez-Alvarez et al., 2017). A large number of descriptions are simply neutral, not to mention negative. To reveal the true worth of an article, one needs a positive evaluation by the academic community. Even more problematic for this approach is the time lag between publishing and citations gathering momentum.

With the availability of full-text articles, citances as a method of identifying breakthroughs has gradually attracted the attention of researchers. Scholars have also

attempted to identify transformative scientific findings by combining citation analysis with content characteristics. Citing sentences, also called citances, are sentences from full-text articles that contains one or more references (Nakov et al., 2004). Citing sentences contain additional information not appearing in abstracts (Elkiss et al., 2008) and, therefore, more accurately represent the contribution of an article to scientific development because the practice of citing sentences is collectively deemed to signify the importance of an article by peer researchers (Radev & Abu-Jbara, 2012). Some scholars have accordingly summarized the contributions of research based on the practice of citing sentences (Chen & Zhuge, 2014).

The peculiar property of citances provides a new direction for recognizing breakthroughs. For example, Guo et al. (2014) combined the citation analysis (the analysis reference duration and highly cited/co-citation) and citance analysis to identify milestones articles that form the “academic chain”. In the citance analysis, the authors used “first”, “broken”, “breakthrough” and other iconic filtered comment words to select the papers. Small et al. (2017) extracted citances with the cue word “\*discover\*” and corresponding references to form “discovery citance-reference” pairs, and subsequently manually screened articles with at least 20 “discovery citances” (293 articles) to identify scientific discoveries (128 articles). In their 2017 study, Small et al. illustrate the important role cue words play in identifying transformative research. A common feature of the aforementioned studies is their use of citances to identify the intrinsic value of references, making up for the limitations in accuracy based on external indicators.

One consistent factor in all the research reviewed above—whether in the form of identifying breakthroughs based on external indicators, content analysis, or a combination of both—is the central role played by the citation relationship. Lacking in these researches is the evaluation by the authors themselves regarding their own research. Whether such an evaluation can be found based on the abstract text is an interesting question to investigate. Research into abstract-based literature classification suggests that it is possible to identify breakthroughs based on the content of abstracts.

In this study, we aim to combine the evaluation of others with self-evaluations to identify potential breakthrough publications. Small et al.’s study (2017) offers insights into breakthrough identification based on linguistic features. However, in this study, the focus was only on the word “discover” and its variants. It makes sense to identify and explore other words that represent an innovation or breakthrough evaluation.

There are two main tasks, then, that we focus on in our study. The first is to search for more words that indicate breakthrough research through word frequency analysis. The second is to identify potential breakthrough publications through the others-self evaluation processes and with the help of classification algorithms.

## Materials and methods

### High-quality breakthrough papers

The first challenge of predicting breakthrough publications is defining a core set of high-quality breakthrough publications to explore how they are cited by others as well as how their authors evaluate the breakthroughs. Due to the explosive growth of publications in the biomedical field, it is difficult to identify breakthrough papers to create a high-quality

data set. In this study, we used articles recognized by the Nobel Prize Committee in physiology or medicine or a *Science* Breakthrough of the Year Award as “ground truths” of biomedical breakthroughs. Each year the Nobel Prize Committee acknowledges the key publications of the prize winners that represent their award-winning achievements and highlight the scientific breakthrough they are being recognized for. Although there is typically a significant time lag between the actual research and it winning a Nobel Prize, the publications nevertheless undeniably represent scientific breakthroughs. We also included those papers recognized with a Science Breakthrough of the Year award. Each year *Science*’s editors and writers choose a significant development as the Breakthrough of the Year (with nine runners-up) and provide the references that resulted in this recognition. The high-quality data set of breakthrough papers, therefore, includes the following:

1. Key publications of Nobel Prize winners in Physiology or Medicine from 1981 and 2018 (for a total of 103 articles); and
2. Publications acknowledged in the Science Breakthrough of the Year award in the biomedical field from 1996 and 2018 (for a total of 556 articles).

Using these two sources, we identified 648 unique breakthrough publications indexed in the PubMed database.

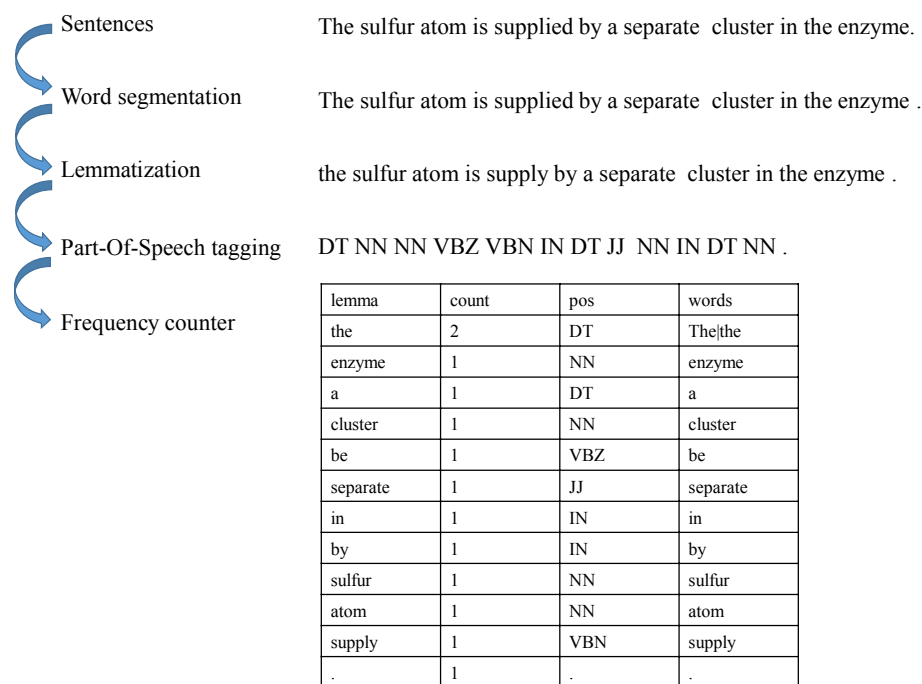
## Breakthrough cue word extraction

Citances are evaluations by others, made by researchers who have read a paper and wish to convey or build upon its ideas in their own work. Abstracts are self-evaluations. They are an author’s summary of their own research, including its benefits and worth. To find the different cue words signaling a possible breakthrough in others- versus self-evaluations, we collected the citances and abstracts of high-quality breakthrough papers and performed both word frequency statistics and manual screening. Using the two databases, Comments on Literature in Literature (Fujiwara & Yamamoto, 2015) and PubMed, we retrieved 135,526 citances and 467 abstracts tied to the high-quality breakthrough papers. The reason why the number of abstracts differs from the number of total breakthrough publications is that some articles do not have abstracts, such as letter.

To identify and extract the cue words, we used the Stanford CoreNLP tool to perform word segmentation and calculate the word frequency statistics for each corpus. The steps are shown in Fig. 1. For each sentence, we first performed word segmentation to reduce the sentences to single words, merging different forms of the same word appearing in the text. Part of speech tagging means to tag all the words according to their context. Word frequencies were calculated after they were lemmatized and tagged. Finally, the cue words were selected manually from a high-frequency word list based on whether the word’s meaning could reflect a breakthrough evaluation.

In the field of information retrieval, Recall and Precision (Cleverdon, 1967) are important indicators of a process’s effectiveness. Hence, these were the metrics we used to evaluate the cue words extracted through our process.

The test dataset included two kinds of papers from the Faculty Opinions (formerly called the F1000) database: those that were designated by at least five reviewers as having a “new



**Fig. 1** The example of word frequency statistics method flow

finding”, i.e., articles that presented novel methods, models, etc., deemed to be breakthroughs, and those that were designated by reviewers as only reflecting “negative/null results”, i.e., articles that presented less-valuable results, regarded as non-breakthroughs. In terms of counts, the test dataset comprised 183 abstracts and 1895 citances from “new finding” publications and 125 abstracts and 1840 citances from “negative/null results” publications.

Recall and Precision were calculated according to Eqs. (1) and (2), respectively. The definitions of abbreviations in the formulas are shown in Table 1. Both indicators reflect the ability to retrieve breakthrough publications using the extracted cue words.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

**Table 1** Definitions of TP, FP, TN, and FN in the Precision and Recall calculation formulas

	Breakthroughs	Non-breakthroughs
Retrieved	TP (true positive)	FN (false negative)
Not retrieved	FP (false positive)	TN (true negative)

## Identifying potential breakthroughs

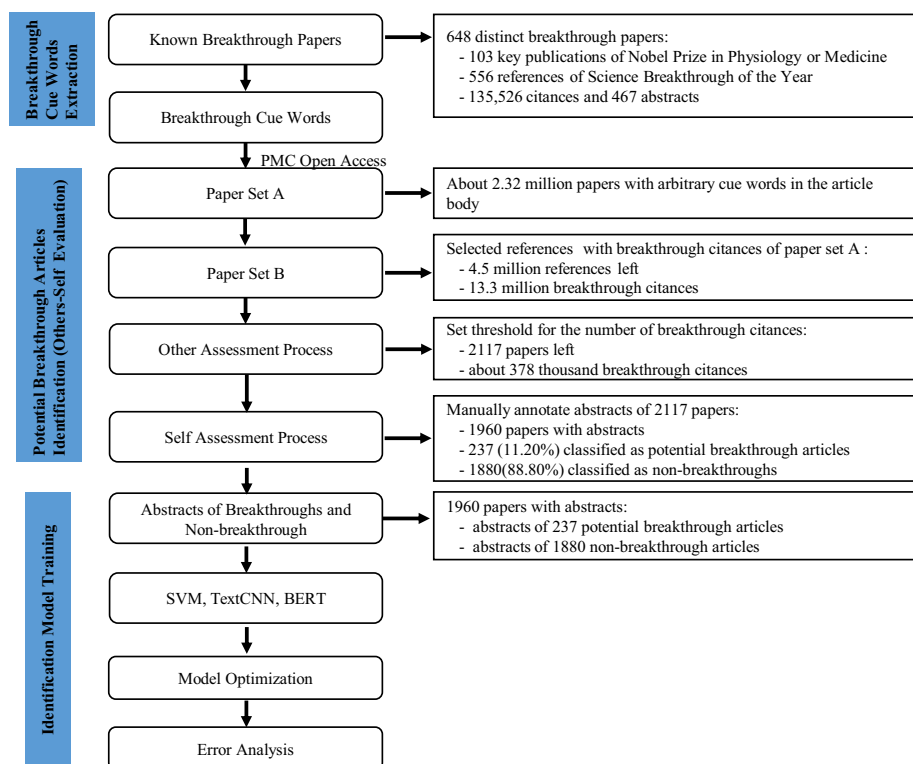
The results in “[Breakthrough cue words](#)” section show that the cue word Precision for the citances was higher than for the abstracts. Hence, to identify breakthrough publications as accurately as possible from such a large volume of literature, we read the full articles with the citances to try and glean some insights into cue word selection. What we found was that the cue words were largely used by the author to describe the work in references, such as in a Literature Review. Obviously, our next step was to download all the references (from the CoLIL database) that had been described with a breakthrough keyword and read those articles, too. We developed the lingo of calling a citance containing a breakthrough cue word as a “breakthrough citance”. According to Small et al. (2017), however, a strategy of limiting cue words to those appearing in citances can still fail to find true links between cue words and references or, conversely, create a false association if the reference and cue words occur in the same sentence but are semantically unrelated. These problems can be compensated for (to some extent) by requiring that cue words and specific references occur in multiple citances. Therefore, we selected the papers with at least 100 breakthrough citances for subsequent analysis. Selecting papers with breakthrough citances meeting a threshold is part of an “others-evaluation” process, which means that many researchers use cue words in their evaluation of the papers.

Turning to the process of self-evaluation, where authors provide cue words about the worth of their own research, it was important to consider that abstracts serve many purposes. They can provide a summary of the research, state shortcomings in the literature or outline problems to be solved. They may discuss the intended audience, and they can provide an evaluation of the research or its implications. However, sometimes an abstract may not include a positive evaluation. After all, many scholars tend to err on the side of conservatism and caution, letting others judge the significance of their work. Alternatively, the results may be neutral or negative. Both can make it difficult to determine whether the research is of groundbreaking significance. Further, reviews, surveys, guidelines, statistical reports, etc., might also contain cue words yet not reflect original innovations or major breakthroughs in themselves. Both possibilities made manual screening necessary. During this process, we coded abstracts without positive evaluations plus reviews and the like as “0”, i.e., non-breakthrough. To be coded with a “1”, the article had to meet two criteria: (1) the abstract had to include a definitively positive evaluation; (2) the research results had to include new findings, change prevailing thinking, or prove a scientific phenomenon for the first time. We ultimately implemented a self-assessment process on the approximately 2000 papers selected after the others-evaluation process.

With these manual reviews done, we then used a text classification algorithm to separate the valuable information from the not-so-valuable. We tested three algorithms—SVM (Chang & Lin, 2011), TextCNN (Zhang & Wallace, 2015), and BERT (Devlin et al., 2018)—and chose the best results. Abstracts are highly accessible, so we formed a dataset of abstracts based on papers with 100 or more “breakthrough citances”. 80% of the data was used for training (selected at random), with the remaining 20% used as the test set. The model produced the best result in terms of Precision, Recall, F1-score, and Accuracy was selected for subsequent error analysis.

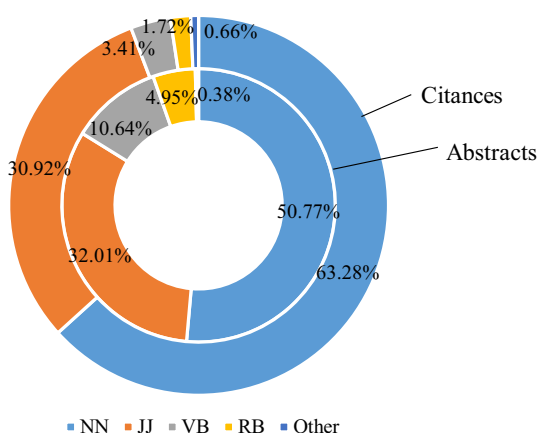
## Results

This section presents the experimental results for each of the steps in the process. A full set of results can be found in Figs. 2, 3, 4, 5, 6 and 7.



**Fig. 2** Biomedical breakthrough papers identification steps and corresponding results

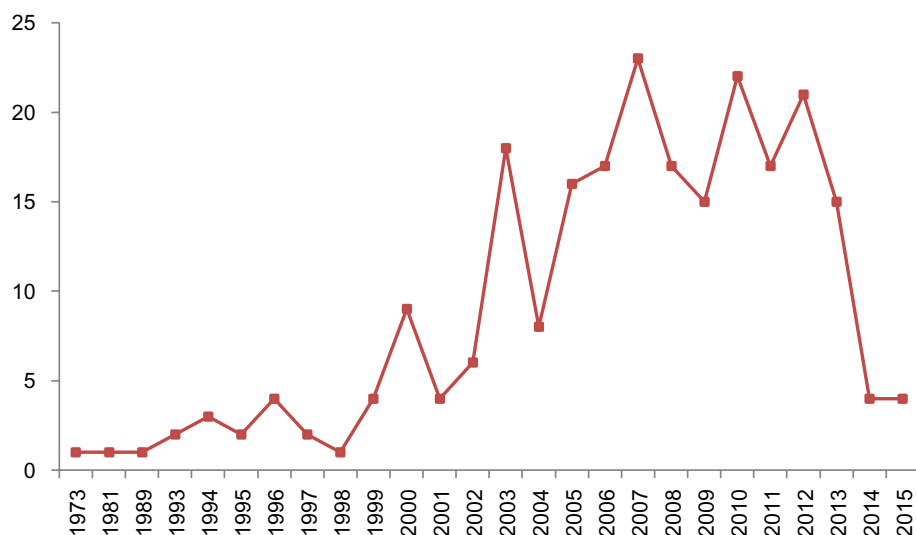
**Fig. 3** The proportion of parts of speech in the abstracts and citations corpora. NN, JJ, VB, RB refer to nouns, adjectives, verbs, and adverbs, respectively



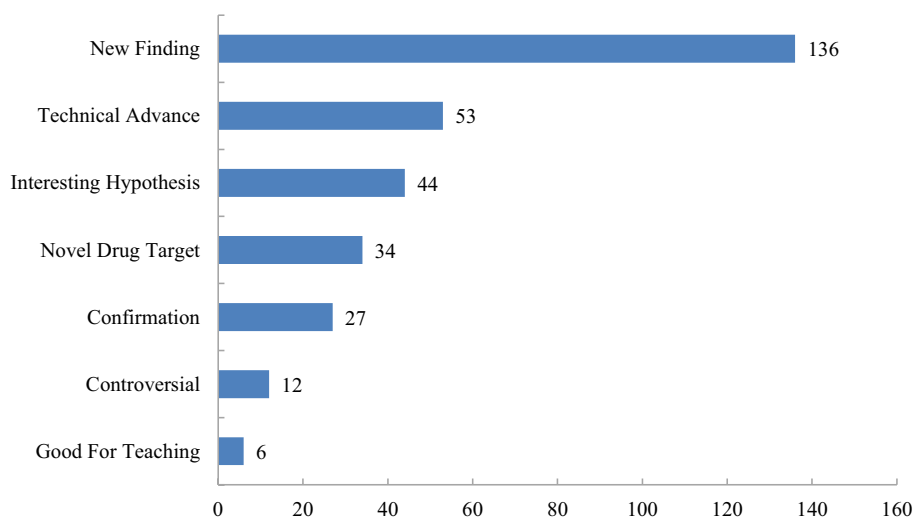
## Breakthrough cue words

After performing word frequency analysis on the abstracts and citations, we extracted 7058 possible cue words from the abstracts and 70,995 from the citations. We merged words





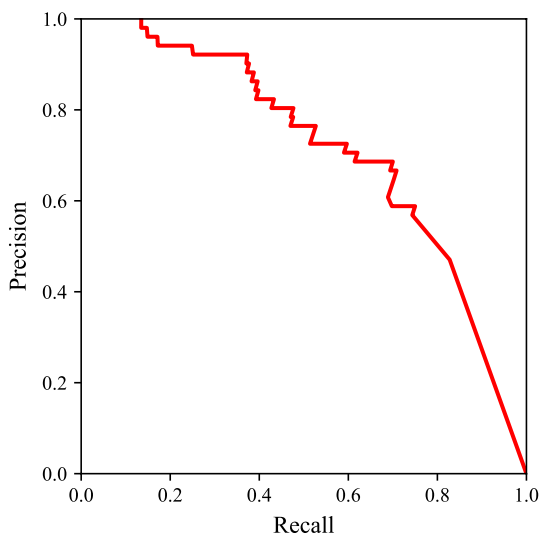
**Fig. 4** The distribution of potential breakthrough papers over time



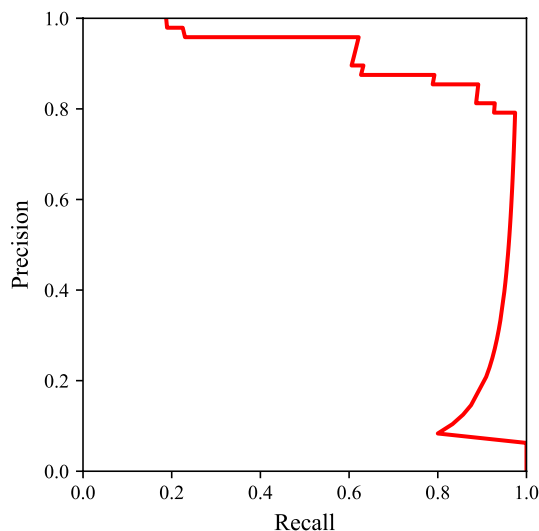
**Fig. 5** The categories of 273 potential breakthrough articles recommended with F1000 and corresponding articles number

with the same stem together and then classified each word as a part of speech, as shown in Fig. 3. Three types of parts of speech dominated the results for both the abstracts and the citances—nouns (NN), adjectives (JJ), and verbs (VB)—indicating that abstracts and citances are similar in composition. However, there were slight differences in the proportions of the three groups. For example, the proportion of verbs in the abstracts was significantly higher than in the citances, suggesting that authors use verbs more frequently when describing their own research.

**Fig. 6** The area under precision-recall curve of BERT model



**Fig. 7** The area under precision-recall curve of BERT-KeyPos-Last sentence model



During the cue word selection process, we took the top 500 most-frequently-used words as candidate terms. Eliminating medical specialty words, like MeSH terms, as well as words unrelated to breakthrough evaluation, we ultimately selected 8 cue words from the abstracts corpus: “new”, “novel”, “potential”, “key”, “change”, “evidence”, “basis”, and “base”, and 8 cue words from the citances corpus: “change”, “first”, “potential”, “new”, “novel”, “since”, “discovery” and “discover”. Table 2 provides the descriptive statistics for the words in these sets. On the whole, whether from the perspective of self-evaluation or others-evaluation, researchers often use the words “new”, “potential”, and “novel” (and their variants) to describe innovative and valuable research in the abstracts. However, further into an article, they tend to use a richer vocabulary to describe research, and their attitudes toward that research tend to be more spelled out more plainly.

**Table 2** The breakthrough cue words selected from the abstracts and citances corpus and the corresponding parts of speech, frequency and original words

Lemma	Source	Word frequency	Part of speech	Original words
Change	Abstracts	87	NN	Changes, change, changed, changes
New	Abstracts	82	JJ	New
Potential	Abstracts	62	JJ	Potential, potential, potentials
Evidence	Abstracts	62	NN	Evidence, evidence, evidenced
Base	Abstracts	53	NN	Based, base, based
Novel	Abstracts	42	JJ	Novel, novel
Key	Abstracts	36	JJ	Key, key
Basis	Abstracts	29	NN	Bases, basis
First	Citances	6145	JJ	First, first
Potential	Citances	4629	NN	Potential, potential, potentials, potentiality, potentialize
Change	Citances	4176	NN	Changes, Changing, change, changed, changes, changing
New	Citances	3874	JJ	New, new
Since	Citances	3716	IN	Since, since
Discovery	Citances	2677	NN	Discoveries, discovery
Discover	Citances	1619	VB	Discovered, discover, discovered, discovering, discovers
Novel	Citances	1481	JJ	Novel, novel

The next step was to calculate the Recall and Precision scores. The plan was to use cue words to retrieve breakthroughs in the test dataset and calculate the Recall and Precision of cue words at retrieving breakthrough publications according to formulae (1) and (2). Different corpora were retrieved with cue words from corresponding sources, and we used all of the original words rather than lemmas. The Recall and Precision for the citances were much higher than for abstracts (citances: Precision 70.77%, Recall 83.13% vs. abstracts: Precision 58.54%, Recall 52.46%). These results indicate that citances cue words are much more effective at retrieving breakthrough publications. Therefore, for the remainder of the study, we only used the citance cue words and dataset for analysis.

### Potential breakthroughs identified via the others-self assessment process

From the PMC open access subset, we retrieved a total of about 2.32 million articles containing at least one cue word somewhere in the full text of the paper. Tracing the references cited in these papers using PubMed resulted in 12 million articles. All citances were downloaded using the unique “PubMed ID” of the reference through the CoLiL database. All data were stored in a MySQL database for subsequent queries and analysis. After eliminating those publications without citances and those without breakthrough citances, roughly 4.5 million articles remained with a total of 13.3 million breakthrough citances (about 3 citances per paper).

As described in the “[Methods](#)” section, we set the threshold of breakthrough citances per paper to 100 in the process of others-evaluation. Only 2117 articles met this threshold. These articles had a total of around 378,000 breakthrough citances, accounting for 2.84% of the total number of citances, with an average of 179 breakthrough citances per article. Of these 2117 articles, the paper *Global Cancer Statistics* (2011) published in 2011, had the highest number of breakthrough citances about 1891. Authors often cite facts taken from statistical reports in the introduction or background sections of their papers, and so these papers tend to accumulate a high citation count over a shorter period of time. But these statistical reports are not breakthrough studies, which shows the necessity of manual screening. In the process of manual screening. After manual screening and, again, removing reviews and the like, only 237 of the 2117 papers met the dual criteria for being classified as a breakthrough candidate. Thus, the majority of papers (88.80%) were classified as non-breakthroughs, which is consistent with our understanding that breakthrough research is quite rare.

After further filtering via the others-self evaluation process, we identified a total of 237 potential breakthrough articles—all published between 1973 and 2015 (Fig. 4). Most were published after 2000, which could be for a variety of reasons. For example, prior to 2000, digitization was rare, and so was open science. Hence, more articles published post-2000 are available for analysis. No breakthrough articles were identified after 2015, which may be related to citation lag and the time it takes to reach 1000 citances.

To further verify the effectiveness of the identification method and determine whether the identified articles did, in fact, represent breakthrough research, we extended our assessment to include the F1000 indicators and recommendation counts per category. Among the 273 candidate papers, 145 had been recommended at least once. Figure 5 shows the categories of the 273 potential breakthrough articles recommended by the F1000. Most articles were recommended as “new finding”; some were considered to be technological advances, interesting hypotheses, or the discovery of novel drug targets.

All the evaluation results prove the effectiveness of the others-self evaluation method in identifying potential breakthroughs. For example, consider the top 20 potential breakthrough articles with the largest number of breakthrough citances (Table 3): 17 articles were recommended as containing new findings, and a few were also designated in the “Novel Drug Target” and “Technical Advance” categories.

## Deep learning results

As identified by the others-self evaluation method, only 237 articles were candidates for breakthrough research and 1880 were non-breakthroughs. Some articles did not contain abstracts, reducing the final training dataset to only 1960 abstracts. We used the SVM, TextCNN, and the BERT algorithms to conduct preliminary training and then optimize the classification models.

Table 4 summarizes the results. From these, we see that all models have a strong ability to predict the samples coded “0”. This may be related to the large difference in the number of breakthrough versus non-breakthroughs samples (the ratio of samples coded was 1 to

**Table 3** The potential breakthrough articles with the largest number of breakthrough citances (Top20) and corresponding counts and categories recommended by F1000

PMID	Publication year	Number of breakthrough citances	Recommendations counts	Recommended as
12,629,218	2003	680	5	New finding; technical advance; interesting hypothesis
17,183,312	2006	425	9	New finding; interesting hypothesis; controversial; technical advance; confirmation
16,923,388	2006	423	10	New finding; interesting hypothesis; technical advance
15,549,107	2004	417	3	New finding; confirmation; interesting hypothesis
11,102,521	2000	410	0	
18,766,170	2008	406	0	
18,772,396	2008	363	2	New finding; confirmation; novel drug target
17,625,570	2007	357	1	New finding
10,990,547	2000	353	0	
12,068,308	2002	343	5	New finding
22,810,696	2012	331	1	New finding; novel drug target
7,509,044	1994	322	1	New finding
20,129,251	2010	311	1	New finding; novel drug target
23,577,628	2013	301	4	New finding
17,051,156	2006	294	2	New finding
23,446,348	2013	294	3	New finding
20,686,565	2010	294	1	Technical advance
11,679,670	2001	289	4	New finding; technical advance
14,505,575	2003	285	1	New finding
20,393,566	2010	283	1	New finding; technical advance

**Table 4** A summary of classification results for the four models, SVM, TextCNN, BERT, and BERT-KeyPos-Last sentence in terms of indicators Recall, Precision, F1-score, and Accuracy

Model	Label	Precision	Recall	F1	Accuracy
SVM	0	0.8830	0.4428	0.5898	0.4643
	1	0.1402	0.6078	0.2279	
	Macro avg	0.5117	0.5253	0.4089	
TextCNN	0	0.8699	1	0.9304	0.8699
	1	0	0	0	
	Macro avg	0.4349	0.5	0.4652	
BERT	0	0.9574	0.9238	0.9403	0.8980
	1	0.5873	0.7255	0.6491	
	Macro avg	0.7724	0.8246	0.7947	
BERT-KeyPos-Last sentence	0	0.9717	0.9562	0.9639	0.9398
	1	0.7885	0.8542	0.8200	
	Macro avg	0.8801	0.9052	0.8919	

“1” indicates breakthrough research, and “0” indicates non-breakthroughs. BERT-KeyPos-Last sentence is an optimization of the BERT model. The strategy adopted is to include only the sentences with positive keywords and the last sentence of abstracts as input for training

0 is 1:7). However, the BERT model showed the stronger ability to identify breakthrough research and achieved a better balance at identifying positive and negative samples. The F1-score of this model was 0.79 with an accuracy of 0.89—both of which are higher than the other two models.

In the self-evaluation screening experiment, we made decisions based on whether the authors had made a positive evaluation of their own research by extracting “judgment sentences”. During the process, we found that, in addition to the above breakthrough cue words, there were some 2-g phrases, such as new view/insight/direction/avenue, first time/report/demonstration, etc., and some 3-g phrases, such as “open the way”, “narrows the gap”, “challenge the view”, etc., that also represented a positive evaluation. Unlike the polysemy of words, the meaning of phrases is more precise. From an analysis of the judgment sentences, 71.3% contained positive keywords and 82.87% appeared at the end of the abstract.

Given the results of the model comparison, we optimized BERT. We extracted the sentences containing positive keywords and the last sentences of abstracts to construct “judgment sentences”. The positive keywords included cue words obtained by word frequency analysis as well as 2-g and 3-g phrases obtained during the self-evaluation screening process. During the training process, we only gave “judgment sentences” as inputs to the BERT model. From a comparison of the area under precision-recall curve and F1-score of the BERT and Bert-KeyPos-Last Sentence models (as shown in Figs. 6, 7 and Table 4), the latter significantly improved the ability to identify abstracts having breakthrough evaluations (F1 = 0.89).

## Error analysis

During the training experiments, the significant difference between the number of breakthroughs and non-breakthroughs caused the traditional machine learning classification

algorithms SVM and TextCNN to experience difficulties in identifying abstracts with positive descriptions, while BERT—a novel and functional deep learning algorithm—performed much better. After optimization, BERT’s classification ability improved even more. Among the above four classification models, the KeyPos-Last sentence model based on the BERT algorithm has the strongest ability for breakthrough identification, and we used it to perform further error analysis.

There were 7 false negatives and 11 false positives in the identification results of the BERT-KeyPos-Last sentence model on the test dataset. The 7 false negatives (where the manual classification said they were breakthrough papers but the machine did not) concerned: identification of new roles in gut microbiota and unannotated RNAs; the discovery of a novel orthobunyavirus; a potential non-invasive molecular marker for colorectal cancer screening (MiR-92); a new strategy named Drop-seq for quickly separating cells into nanoliter-sized aqueous droplets; and mRNA transcripts. The machine judged them as non-breakthrough research, perhaps because the last sentence of the abstract could not be used as a basis for judgment or because the breakthrough meaning of keywords was not obvious. The 11 false positives (where the deep learning said they were breakthroughs but the manual classification said they were not) were primarily because the abstract contained breakthrough keywords but whose meaning was assessed as non-breakthrough during manual screening. Although the deep learning model had errors in identifying the abstracts having breakthrough evaluation, it can still quickly find many high-quality and highly-evaluated articles from a large number of articles and can be of considerable assistance in breakthrough identification.

## Discussion

Breakthrough research can greatly advance scientific progress and development. If a breakthrough can be identified in its early stages, more funding can be invested and invested sooner into these research fields. It may then be that the breakthrough occurs more quickly. Researchers have made exploratory attempts at identifying breakthrough research but, so far, no method has proven particularly worthwhile. With this study, we hope we have presented a promising method for filtering breakthrough research out of academia’s vast knowledge stores. However, we openly admit that early recognition remains elusive.

In this article, we proposed an “others-self” dual evaluation method for breakthrough identification. The method combines evaluations of others’ work via citances and an author’s evaluations of their own research via claims made in abstracts. To assess the efficacy of the method, we compared our results to the F1000 insights, derived through qualitative evaluations of research. The results showed that most of the articles we identified are included in the F1000 database, recommended in one or more of the categories New Finding, Novel Drug Target and/or Technical Advance. Hence, we are confident in claiming that the others-self evaluation method can identify high-value publications. Additionally, we input the abstracts to the deep learning model in the hopes that it could automatically identify ones with positive evaluations, which simplifies the process of self evaluation by replacing manually review each article. For interest’s sake, the results showed that the BERT-KeyPos-Last sentence model has the best identification ability with an F1 score of 0.89. However, there are some process issues that need to be resolved before we can deem the method truly useful—the main being the influence of citation lag. During our process, we had to set the threshold for the number of breakthrough citances to 100 to distill the

candidate articles down to a manageable number. As a result, all recent publications were eliminated from the candidate list, leaving the last article at near to six years old. More on these points is discussed alongside some other limitations of the study shortly.

The “others-self” dual evaluation process can also be regarded as a way of evaluating articles. When evaluating a paper, the authors’ own evaluation and the assessment of authors who cite an article are taken into account, which enriches traditional academic evaluation methods. In traditional academic evaluations, the number of citations is still the main index that reflects the academic value and influence of an article. But not all citations are equivalent: some are positive, and most are neutral, and some are even negative (Radev & Abu-Jbara, 2012). During the others-evaluation process, we used the number of citances containing breakthrough cue words to evaluate each article on the basis that the more breakthrough citances, the greater the academic value of the article. Although this method of evaluating articles evaluation is more accurate, it has limited usefulness for recent publications due to citation lag (see Fig. 4).

However, for recent publications or articles with a relatively small number of citations, the automatic identification model tested in this study can be used to classify abstracts and identify significant articles from the author’s own perspective for the purpose of article evaluation. It is not uncommon for breakthrough research to be overlooked or ignored for years, or for it to conflict with existing scientific paradigms and therefore be harshly critiqued by the scientific community. Over time, these “sleeping beauty” publications are awakened by a “prince” publication and so ultimately accumulate a large number of citations. This is a self-reinforcing loop where citations confer importance to the research, perpetuating more approval and more citations. However, authors often have a global overview of their research area when publishing their papers, so they are well aware of the significance of their findings and often indicate as much in their abstract. Abstract-based identification of breakthroughs is possible and provides a new avenue for evaluating recent publications or special articles.

The method proposed in this study, therefore, has practical applications in several respects. This dual evaluation method can be used to accurately identify the potential breakthrough articles in any biomedical subfield by first filtering the number of breakthrough citances and then adopting the model to identify those abstracts having positive evaluations. For recent publications or those with few citations, the approach can directly identify abstracts with positive evaluations and offer a point of departure for subsequent study and analysis.

In our research, we also attempted to explore how authors who cite breakthrough articles differ in their description and evaluation compared to those made by the actual author of the breakthrough article. It turns out that authors who employ citations often use nouns to describe references. Although the authors of breakthrough papers also used nouns to describe their own research, the proportion of verbs increased significantly, which likely reflects their involvement in the research. From the perspective of the academic community, authors who cite breakthrough papers often use “change”, “first”, “potential”, “new”, “novel”, “since”, “discovery”, and “discover” to evaluate breakthrough articles. From the perspective of the authors of breakthrough articles, they prefer to use “new”, “novel”, “potential”, “key”, “change”, “evidence”, “basis”, and “base” to describe the significance of their own research. We also explored whether these words can be used to retrieve breakthrough research. As it turns out, both Recall and Precision of retrieval based on citance cue words are much higher than that based on cue words in abstracts. Citance-based retrieval is likely more effective at identifying valuable and significant research compared with the more commonly used abstract field search because the knowledge and contribution



mentioned in citances are what the peers think has an important influence on their research. Therefore, retrieval based on citances may be more helpful to find breakthrough articles.

There are also several limitations to our study. First, this method may not be suitable for recent publications or articles with few citations. Even though the threshold for the number of breakthrough citances per article is adjustable, we still used a threshold during the others-evaluation process to narrow the number of articles, and therefore potentially excluded articles with breakthrough potential. In the process of extracting breakthrough text features from abstracts and citances, we also selected breakthrough cue words from the top 500 words with the highest frequency, ignoring the words that appear less frequently but represent strong positive emotions, such as “milestones” and “landmarks” (even the term “breakthrough” was excluded due to its low frequency). Another limitation is that we regarded the citances containing breakthrough cue words as breakthrough citances during the process of “others-evaluation”. But, because of polysemy, the same word can be used in different contexts. For example, “new case”, “nucleic acid base”, etc., do not indicate breakthrough evaluations but are commonly used phrases in the biomedical field. Therefore, it is challenging to determine whether a citance containing breakthrough cue words has a breakthrough meaning, which leads to the unreliability of screening results by “others-evaluation” (although we compensated for this limitation by creating another by setting a relatively high threshold for the number of breakthrough citances).

There remain significant future opportunities for research into breakthrough identification. Most current studies on breakthroughs (including this one) remain focused on retrospective identification. However, early identification is more meaningful as it could provide useful guidance for scientific research planning and funding, as well as a new research direction for researchers. In the above experiments, we have shown the feasibility of using a text classification algorithm to identify breakthrough articles. Although it remains more practical to identify valuable articles based on whether there is a breakthrough evaluation in the abstract (because almost all abstracts are openly accessible in the literature database), identifying them through citances is more meaningful because they represent (in concentrated form) the value the academic community places on important original research. At the same time, citances have the power to signal breakthrough research early. While an article with a low citation count does not suggest it has the potential to be a breakthrough article, if its citances repeatedly show the article contains a “first finding...”, it strongly suggests that the article may have groundbreaking potential.

At present, researchers have studied the polarity of citation, dividing into positive, negative, and neutral (Abu-Jbara et al., 2013), or identifying meaningful citations from all citations (Hassan et al., 2018; Valenzuela et al., 2015). Alternatively, we suggest using text classification algorithms to divide citances into those with and without breakthrough evaluations. This method can be used to identify recent publications or articles with breakthrough potential despite their relatively small number of citations. It is worth noting that the results obtained by these identification methods are only references. Whether these studies are truly groundbreaking remains an open question, and recent examples of articles with important findings that are later retracted for academic misconduct give us pause. These issues cannot be resolved in advance, making identification of breakthrough research or major scientific discoveries more difficult.

**Acknowledgements** This work was supported by CAMS Initiative for Innovative Medicine (2016-I2M-3-018), and NSTL “Key Technology Optimization Integration and System Development of Next Generation Open Knowledge Service Platform” (2020XM05).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 2013* (pp. 596–606): Association for Computational Linguistics.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–27. <https://doi.org/10.1145/1961189.1961199>
- Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3), 431–449. <https://doi.org/10.1002/asi.21694>
- 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. <https://doi.org/10.1016/j.joi.2009.03.004>
- Chen, J. Q., & Zhuge, H. (2014). Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems-the International Journal of Esience*, 32, 246–252. <https://doi.org/10.1016/j.future.2013.07.018>
- Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6), 173–194. <https://doi.org/10.1108/eb050097>
- Comins, J. A., & Hussey, T. W. (2015). Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy. *Scientometrics*, 104(2), 575–580. <https://doi.org/10.1007/s11192-015-1598-2>
- Comins, J. A., & Leydesdorff, L. (2017). Citation algorithms for identifying research milestones driving biomedical innovation. *Scientometrics*, 110(3), 1495–1504. <https://doi.org/10.1007/s11192-016-2238-1>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv e-prints, <http://arxiv.org/abs/1810.04805>
- Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1), 51–62. <https://doi.org/10.1002/asi.20707>
- Fujiwara, T., & Yamamoto, Y. (2015). CoLiL: a database and search service for citation contexts in the life sciences domain. *Journal of Biomedical Semantics*, 6, 38. <https://doi.org/10.1186/s13326-015-0037-x>
- Garfield, E., & Welljams-Dorof, A. (1992). Of Nobel class: A citation perspective on high impact research authors. *Theoretical Medicine*, 13(2), 117–135. <https://doi.org/10.1007/bf02163625>
- Guo, Q., Du, J., & Tang, X. A Bibliometric Framework for Identifying “Academic Chain” —A Case Study of 2014 Nobel Prize for Chemistry. In *Proceedings of the 23rd international conference on science and technology indicators*, 2018-09-11 (pp. 331–338): Centre for Science and Technology Studies (CWTS).
- Hassan, S.-U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics*, 117(3), 1645–1662. <https://doi.org/10.1007/s11192-018-2944-y>
- Hernandez-Alvarez, M., Soriano, J. M. G., & Martinez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4), 561–588. <https://doi.org/10.1017/S1351324916000346>
- Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. In C. Hannaway (Ed.), *Biomedicine in the Twentieth Century: Practices, Policies, and Politics, volume 72 of Biomedical and Health Research*. (pp. 317–353). National Institutes of Health.

- Huang, Y. H., Ko, M. T., Hsu, C. N. (2014). Identifying transformative research in biomedical sciences. In: S. M. Cheng, & M. Y. Day (Eds.), *Technologies and applications of artificial intelligence*. TAAI 2014. Lecture Notes in Computer Science, vol. 8916. Springer, Cham. [https://doi.org/10.1007/978-3-319-13987-6\\_18](https://doi.org/10.1007/978-3-319-13987-6_18)
- Huang, Y., Hsu, C. & Lerman, K. (2013). Identifying transformative scientific research. In *2013 IEEE 13th international conference on data mining, Dallas, TX, USA* (pp. 291–300). <https://doi.org/10.1109/ICDM.2013.120>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Nakov, P. I., Schwartz, A. S., & Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR '04 workshop on Search and Discovery in Bioinformatics*. National Institutes of Health. NIH Director's Transformative Research Award. (2016). Retrieved December 23, 2019 from <https://commonfund.nih.gov/tra/description>
- National Science Board. (2007). Enhancing Support of Transformative Research at the National Science Foundation. Retrieved December 24, 2019 from [https://www.nsf.gov/nsb/documents/2007/tr\\_report.pdf](https://www.nsf.gov/nsb/documents/2007/tr_report.pdf)
- National Science Foundation. (2015). Definition of Transformative Research. Retrieved December 24, 2019 from [https://www.nsf.gov/about/transformative\\_research/definition.jsp](https://www.nsf.gov/about/transformative_research/definition.jsp)
- Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81, 49–55. <https://doi.org/10.1016/j.techfore.2012.09.017>
- Radev, D., & Abu-Jbara, A. (2012). Rediscovering ACL discoveries through the lens of ACL anthology network citing sentences. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju Island, Korea, jul 2012* (pp. 1–12): Association for Computational Linguistics
- Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, 11(1), 46–62. <https://doi.org/10.1016/j.joi.2016.11.001>
- Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying Meaningful Citations. Paper presented at the AAAI Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence
- Winnink, J. J., Tijssen, R. J. W., & van Raan, A. F. J. (2016). Theory-changing breakthroughs in science: The impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology*, 67(5), 1210–1223. <https://doi.org/10.1002/asi.23505>
- Winnink, J. J., Tijssen, R. J. W., & van Raan, A. F. J. (2019). Searching for new breakthroughs in science: How effective are computerised detection algorithms? *Technological Forecasting and Social Change*, 146, 673–686. <https://doi.org/10.1016/j.techfore.2018.05.018>
- Wolcott, H. N., Fouch, M. J., Hsu, E. R., DiJoseph, L. G., Bernaciak, C. A., Corrigan, J. G., et al. (2016). Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers. *Scientometrics*, 107(2), 807–817. <https://doi.org/10.1007/s11192-016-1861-1>
- Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv e-prints, <http://arxiv.org/abs/1510.03820>
- 19th National Congress of the Communist Party of China. (2017). Secure a decisive victory in building a moderately prosperous society in all respects and strive for the great success of socialism with Chinese characteristics for a New Era. Retrieved December 03, 2019 from <http://www.china.org.cn/20171105-002.pdf>