



RESEARCH ARTICLE

scite: A smart citation index that displays the context of citations and classifies their intent using deep learning

Josh M. Nicholson¹ , Milo Mordaunt¹ , Patrice Lopez² , Ashish Uppala¹ ,
Domenic Rosati¹ , Neves P. Rodrigues¹ , Peter Grabitz^{1,3} , and Sean C. Rife^{1,4} 

¹scite, Brooklyn, NY, USA

²science-miner, France

³Charite Universitaetsmedizin Berlin, Berlin, Germany

⁴Murray State University, Murray, KY, USA

an open access  journal



Citation: Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., & Rife, S. C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3), 882–898. https://doi.org/10.1162/qss_a_00146

DOI:
https://doi.org/10.1162/qss_a_00146

Peer Review:
https://publons.com/publon/10.1162/qss_a_00146

Received: 15 March 2021
Accepted: 29 June 2021

Corresponding Author:
Josh M. Nicholson
josh@scite.ai

Handling Editor:
Ludo Waltman

Copyright: © 2021 Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife.
Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



Keywords: bibliometrics, citations, evaluation, machine learning, publishing, scientometrics

ABSTRACT

Citation indices are tools used by the academic community for research and research evaluation that aggregate scientific literature output and measure impact by collating citation counts. Citation indices help measure the interconnections between scientific papers but fall short because they fail to communicate contextual information about a citation. The use of citations in research evaluation without consideration of context can be problematic because a citation that presents contrasting evidence to a paper is treated the same as a citation that presents supporting evidence. To solve this problem, we have used machine learning, traditional document ingestion methods, and a network of researchers to develop a “smart citation index” called *scite*, which categorizes citations based on context. Scite shows how a citation was used by displaying the surrounding textual context from the citing paper and a classification from our deep learning model that indicates whether the statement provides supporting or contrasting evidence for a referenced work, or simply mentions it. Scite has been developed by analyzing over 25 million full-text scientific articles and currently has a database of more than 880 million classified citation statements. Here we describe how scite works and how it can be used to further research and research evaluation.

1. INTRODUCTION

Citations are a critical component of scientific publishing, linking research findings across time. The first citation index in science, created in 1960 by Eugene Garfield and the Institute for Scientific Information, aimed to “be a spur to many new scientific discoveries in the service of mankind” (Garfield, 1959). Citation indices have facilitated the discovery and evaluation of scientific findings across all fields of research. Citation indices have also led to the establishment of new research fields, such as bibliometrics, scientometrics, and quantitative studies, which have been informative in better understanding science as an enterprise. From these fields have come a variety of citation-based metrics, such as the *h*-index, a measurement of researcher impact (Hirsch, 2005); the Journal Impact Factor (JIF), a measurement of journal impact (Garfield, 1955, 1972); and the citation count, a measurement of article impact. Despite the

widespread use of bibliometrics, there have been few improvements in citations and citation indices themselves. Such stagnation is partly because citations and publications are largely behind paywalls, making it exceedingly difficult and prohibitively expensive to introduce new innovations in citations or citation indices. This trend is changing, however, with open access publications becoming the standard (Piwowar, Priem, & Orr, 2019) and organizations such as the Initiative for Open Citations (Initiative for Open Citations, 2017; Peroni & Shotton, 2020) helping to make citations open. Additionally, with millions of documents being published each year, creating a citation index is a large-scale challenge involving significant financial and computational costs.

Historically, citation indices have only shown the connections between scientific papers without any further contextual information, such as why a citation was made. Because of the lack of context and limited metadata available beyond paper titles, authors, and the date of publications, it has only been possible to calculate how many times a work has been cited, not analyze broadly *how* it has been cited. This is problematic given citations' central role in the evaluation of research. In short, not all citations are made equally, yet we have been limited to treating them as such.

Here we describe scite (scite.ai), a new citation index and tool that takes advantage of recent advances in artificial intelligence to produce "Smart Citations." Smart Citations reveal how a scientific paper has been cited by providing the context of the citation and a classification system describing whether it provides supporting or contrasting evidence for the cited claim, or if it just mentions it.

Such enriched citation information is more informative than a traditional citation index. For example, when Viganó, von Schubert et al. (2018) cites Nicholson, Macedo et al. (2015), traditional citation indices report this citation by displaying the title of the citing paper and other bibliographic information, such as the journal, year published, and other metadata. Traditional citation indices do not have the capacity to examine contextual information or how the citing paper used the citation, such as whether it was made to support or contrast the findings of the cited paper or if it was made in the introduction or the discussion section of the citing paper. Smart Citations display the same bibliographical information shown in traditional citation indices while providing additional contextual information, such as the citation statement (the sentence containing the in-text citation from the citing article), the citation context (the sentences before and after the citation statement), the location of the citation within the citing article (Introduction, Materials and Methods, Results, Discussion, etc.), the citation type indicating intent (supporting, contrasting, or mentioning), and editorial information from Crossref and PubMed, such as corrections and whether the article has been retracted (Figure 1). Scite previously relied on Retraction Watch data but moved away from this due to licensing issues. Going forward, scite will use its own approach¹ to retraction detection, as well as data from Crossref and PubMed.

Adding such information to citation indices has been proposed before. In 1964, Garfield described an "intelligent machine" to produce "citation markers," such as "critique" or, jokingly, "calamity for mankind" (Garfield, 1964). Citation types describing various uses of citations have been systematically described by Peroni and Shotton in CiTO, the *Citation Typing Ontology* (Peroni & Shotton, 2012). Researchers have used these classifications or variations of them in several bibliometric studies, such as the analysis of citations (Suelzer, Deal et al., 2019) made to

¹ Details of how retractions and other editorial notices can be detected through an automated examination of metadata—even when there is no explicit indication that such notice(s) exist—will be made public via a manuscript currently in preparation.

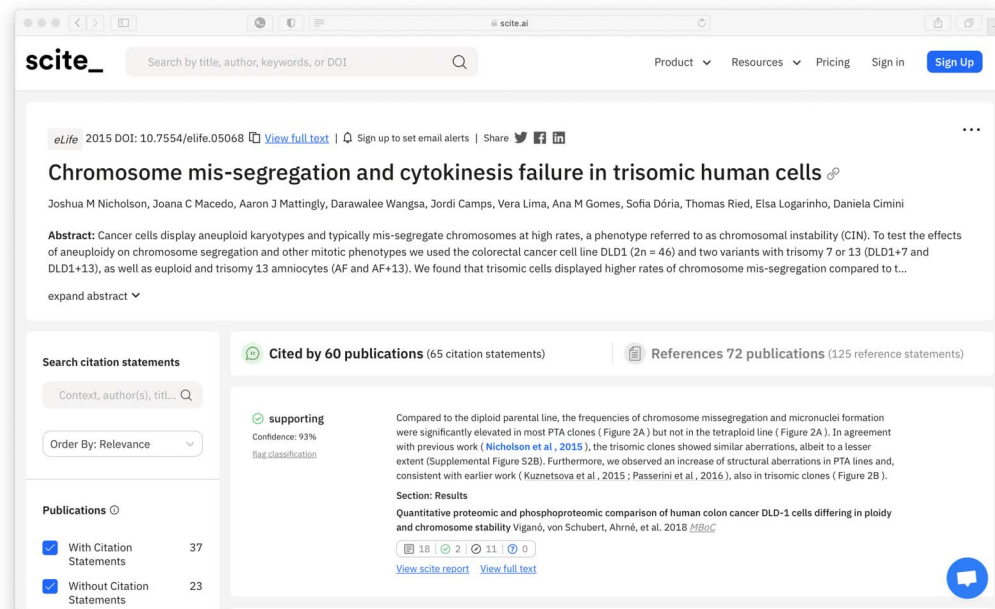


Figure 1. Example of scite report page. The scite report page shows citation context, citation type, and various features used to filter and organize this information, including the section where the citation appears in the citing paper, whether or not the citation is a self-citation, and the year of the publication. The example scite report shown in the figure can be accessed at the following link: <https://scite.ai/reports/10.7554/elife.05068>.

the retracted Wakefield paper (Wakefield, Murch et al., 1998), which found most citations to be negative in sentiment. Leung, Macdonald et al. (2017) analyzed the citations made to a five-sentence letter purporting to show opioids as nonaddictive (Porter & Jick, 1980), finding that most citations were uncritically citing the work. Based on these findings, the journal appended a public health warning to the original letter. In addition to citation analyses at the individual article level, citation analyses taking into account the citation type have also been performed on subsets of articles or even entire fields of research. Greenberg (2009) discovered that citations were being distorted, for example being used selectively to exclude contradictory studies to create a false authority in a field of research, a practice carried into grant proposals. Selective citing might be malicious, as suggested in the Greenberg study, but it might also simply reflect sloppy citation practices or citing without reading. Indeed, Letrud and Hernes (2019) recently documented many cases where people were citing reports for the opposite conclusions than the original authors made.

Despite the advantages of citation types, citation classification and analysis require substantial manual effort on the part of researchers to perform even small-scale analyses (Pride, Knoth, & Harag, 2019). Automating the classification of citation types would allow researchers to dramatically expand the scale of citation analyses, thereby allowing researchers to quickly assess large portions of scientific literature. PLOS Labs attempted to enhance citation analysis with the introduction of “rich citations,” which included various additional features to traditional citations such as retraction information and where the citation appeared in the citing paper (PLOS, 2015). However, the project seemed to be mostly a proof of principle, and work on rich citations stopped in 2015, although it is unclear why. Possible reasons that the project did not mature reflect the challenges of accessing the literature at scale, finding a suitable business model for the application, and classifying citation types with the necessary precision

and recall for it to be accepted by users. It is only recently that machine learning techniques have evolved to make this task possible, as we demonstrate here. Additional resources, such as the Colil Database (Fujiwara & Yamamoto, 2015) and SciRide Finder (Volanakis & Krawczyk, 2018) both allow users to see the citation context from open access articles indexed in PubMed Central. However, adoption seems to be low for both tools, presumably due to limited coverage of only open access articles. In addition to the development of such tools to augment citation analysis, various researchers have performed automated citation typing. Machine learning was used in early research to identify citation intent (Teufel, Siddharthan, & Tidhar, 2006) and recently Cohan, Ammar et al. (2019) used deep learning techniques. Athar (2011), Yousif, Niu et al. (2019), and Yan, Chen, and Li (2020) also used machine learning to identify positive and negative sentiments associated with the citation contexts.

Here, by combining the largest citation type analysis performed to date and developing a useful user interface that takes advantage of the extra contextual information available, we introduce scite, a smart citation index.

2. METHOD

2.1. Overview

Smart citations are created by extracting and analyzing citation statements from full-text scientific articles. This process is broken into four major steps (see Figure 2):

1. The retrieval of scientific articles
2. The identification and matching of in-text citations and references within a scientific article
3. The matching of references against a bibliographic database

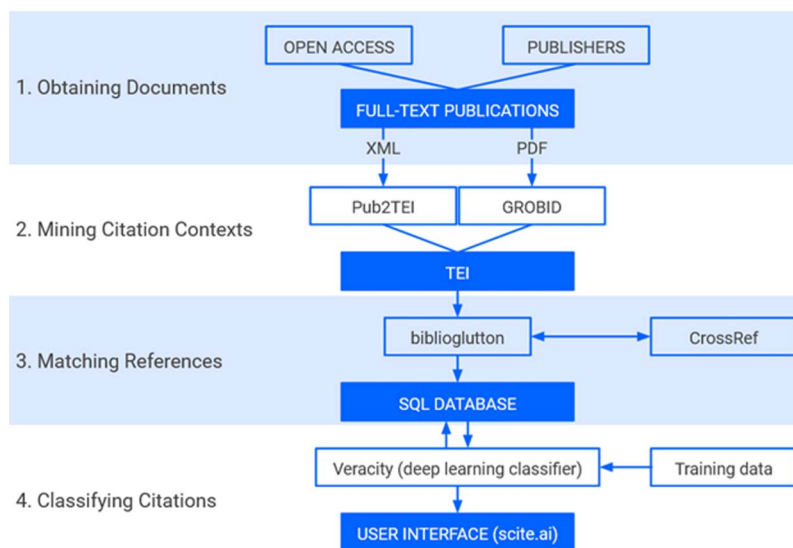


Figure 2. The scite ingestion process. Documents are retrieved from the internet, as well as being received through file transfers directly from publishers and other aggregators. They are then processed to identify citations, which are then tied to items in a paper's reference list. Those citations are then verified, and the information is inserted into scite's database.

4. The classification of the citation statements into citation types using deep learning.

We describe the four components in more detail below.

2.2. Retrieval of Scientific Documents

Access to full-text scientific articles is necessary to extract and classify citation statements and the citation context. We utilize open access repositories such as PubMed Central and a variety of open sources as identified by Unpaywall (Else, 2018), such as open access publishers' websites, university repositories, and preprint repositories, to analyze open access articles. Other relevant open access document sources, such as Crossref TDM and the Internet Archive have been and are continually evaluated as new sources for document ingestion. Subscription articles used in our analyses have been made available through indexing agreements with over a dozen publishers, including Wiley, BMJ, Karger, Sage, Europe PMC, Thieme, Cambridge University Press, Rockefeller University Press, IOP, Microbiology Society, Frontiers, and other smaller publishers. Once a source of publications is established, documents are retrieved on a regular basis as new articles become available to keep the citation record fresh. Depending on the source, documents may be retrieved and processed anywhere between daily and monthly.

2.3. Identification of In-Text Citations and References from PDF and XML Documents

A large majority of scientific articles are only available as PDF files², a format designed for visual layout and printing, not text-mining. To match and extract citation statements from PDFs with high fidelity, an automated process for converting PDF files into reliable structured content is required. Such conversion is challenging, as it requires identifying in-text citations (the numerical or textual callouts that refer to a particular item in the reference list), identifying and parsing the full bibliographical references in the reference list, linking in-text citations to the correct items in this list, and linking these items to their digital object identifiers (DOIs) in a bibliographic database. As our goal is to eventually process all scientific documents, this process must be scalable and affordable. To accomplish this, we utilize GROBID, an open-source PDF-to-XML converter tool for scientific literature (Lopez, 2020a). The goal of GROBID is to automatically convert scholarly PDFs into structured XML representations suitable for large-scale analysis. The structuration process is realized by a cascade of supervised machine learning models. The tool is highly scalable (around five PDF documents per second on a four-core server), is robust, and includes a production-level web API, a Docker image, and benchmarking facilities. GROBID is used by many large scientific information service providers, such as ResearchGate, CERN, and the Internet Archive to support their ingestion and document workflows (Lopez, 2020a). The tool is also used for creating machine-friendly data sets of research papers, for instance, the recent CORD-19 data set (Wang, Lo et al., 2020).

Particularly relevant to scite, GROBID was benchmarked as the best open source bibliographical references parser by Tkaczyk, Collins et al. (2018) and has a relatively unique focus on citation context extraction at scale, as illustrated by its usage for building the large-scale

² As an illustration, the ISTEEX project has been an effort from the French state leading to the purchase of 23 million full text articles from the mainstream publishers (Elsevier, Springer-Nature, Wiley, etc.) mainly published before 2005, corresponding to an investment of €55 million in acquisitions. The delivery of full text XML when available was a contractual requirement, but an XML format with structured body could be delivered by publishers for only around 10% of the publications.

Semantic Scholar Open Research Corpus (S2ORC), a corpus of 380.5 million citations, including citation mentions excerpts from the full-text body (Lo, Wang et al., 2020).

In addition to PDFs, some scientific articles are available as XML files, such as the Journal Article Tag Suite (JATS) format. Formatting articles in PDF and XML has become standard practice for most mainstream publishers. While structured XML can solve many issues that need to be addressed with PDFs, XML full texts appear in a variety of different native publisher XML formats, often incomplete and inconsistent from one to another, loosely constrained, and evolving over time into specific versions.

To standardize the variety of XML formats we receive into a common format, we rely upon the open-source tool Pub2TEI (Lopez, 2020b). Pub2TEI converts various XML styles from publishers to the same standard TEI format as the one produced by GROBID. This centralizes our document processing across PDF and XML sources.

2.4. Matching References Against the Bibliographic Database Crossref

Once we have identified and matched the in-text citation to an item in a paper's reference list, this information must be validated. We use an open-source tool, biblio-glutton (Lopez, 2020c), which takes a raw bibliographical reference, as well as optionally parsed fields (title, author names, etc.) and matches it against the Crossref database—widely regarded as the industry standard source of ground truth for scholarly publications³. The matching accuracy of a raw citation reaches an F-score of 95.4 on a set of 17,015 raw references associated with a DOI, extracted from a data set of 1,943 PMC articles⁴ compiled by Constantin (2014). In an end-to-end perspective, still based on an evaluation with the corpus of 1,943 PMC articles, combining GROBID PDF extraction of citations and bibliographical references with biblio-glutton validations, the pipeline successfully associates around 70% of citation contexts to cited papers with correctly identified DOIs in a given PDF file. When the full-text XML version of an article is available from a publisher, references and linked citation contexts are normally correctly encoded, and the proportion of fully solved citation contexts corresponding to the proportion of cited paper with correctly identified DOIs is around 95% for PMC XML JATS files. The scite platform today only ingests publications with a DOI and only matches references against bibliographical objects with a registered DOI. The given evaluation figures have been calculated relative to these types of citations.

2.5. Task Modeling and Training Data

Extracted citation statements are classified into supporting, contrasting, or mentioning, to identify studies that have tested the claim and to evaluate how a scientific claim has been evaluated in the literature by subsequent research.

We emphasize that scite is not doing sentiment analysis. In natural language processing, sentiment analysis is the study of affective and subjective statements. The most common affective state considered in sentiment analysis is a mere polar view from positive sentiment to negative sentiment, which appeared to be particularly useful in business applications (e.g., product reviews and movie reviews). Following this approach, a subjective polarity can be associated with a citation to try to capture an opinion about the cited paper. The evidence used for sentiment classification relies on the presence of affective words in the citation

³ For more information on the history and prevalence of Crossref, see <https://www.crossref.org/about/>.

⁴ The evaluation data and scripts are available on the project GitHub repository; see biblio-glutton (Lopez, 2020c).

context, with an associated polarity score capturing the strength of the affective state (Athar, 2014; Halevi & Schimming, 2018; Hassan, Imran et al., 2018; Yousif et al., 2019). Yan et al. (2020), for instance, use a generic method called *SenticNet* to identify sentiments in citation contexts extracted from PubMed Central XML files, without particular customization to the scientific domain (only a preprocessing to remove the technical terms from the citation contexts is applied). SenticNet uses a polarity measure associated with 200,000 natural language concepts, propagated to the words and multiword terms realizing these concepts.

In contrast, scite focuses on the authors' reasons for citing a paper. We use a discrete classification into three discursive functions relative to the scientific debate; see Murray, Lamers et al. (2019) for an example of previous work with typing citations based on rhetorical intention. We consider that for capturing the reliability of a claim, a classification decision into supporting or contrasting must be backed by scientific arguments. The evidence involved in our assessment of citation intent is directed to the factual information presented in the citation context, usually statements about experimental facts and reproducibility results or presentation of a theoretical argument against or agreeing with the cited paper.

Examples of supporting, contrasting, and mentioning citation statements are given in Table 1, with explanations describing why they are classified as such, including examples where researchers have expressed confusion or disagreement with our classification.

Importantly, just as it is critical to optimize for accuracy of our deep learning model when classifying citations, it is equally important to make sure that the right terminology is used and understood by researchers. We have undergone multiple iterations of the design and display of citation statements and even the words used to define our citation types, including using previous words such as *refuting* and *disputing* to describe contrasting citations and *confirming* to describe supporting citations. The reasons for these changes reflect user feedback expressing confusion over certain terms as well as our intent to limit any potentially inflammatory interpretations. Indeed, our aim with introducing these citation types is to highlight differences in research findings based on evidence, not opinion. The main challenge of this classification task is the highly imbalanced distribution of the three classes. Based on manual annotations of different publication domains and sources, we estimate the average distribution of citation statements as 92.6% mentioning, 6.5% supporting, and 0.8% contrasting statements. Obviously, the less frequent the class, the more valuable it is. Most of the efforts in the development of our automatic classification system have been directed to address this imbalanced distribution. This task has required first the creation of original training data by experts—scientists with experience in reading and interpreting scholarly papers. Focusing on data quality, the expert classification was realized by multiple-blind manual annotation (at least two annotators working in parallel on the same citation), followed by a reconciliation step where the disagreements were further discussed and analyzed by the annotators. To keep track of the progress of our automatic classification over time, we created a holdout set of 9,708 classified citation records. To maintain a class distribution as close as possible to the actual distribution in current scholarly publications, we extracted the citation contexts from Open Access PDF of Unpaywall by random sampling with a maximum of one context per document.

We separately developed a working set where we tried to oversample the two less frequent classes (supporting, contrasting) with the objective of addressing the difficulties implied by the imbalanced automatic classification. We exploited the classification scores of our existing classifiers to select more likely supporting and contrasting statements for manual classification. At the present time, this set contains 38,925 classified citation records. The automatic classification system was trained with this working set, and continuously evaluated with the

Table 1. Real-world examples of citation statement classifications with examples explaining why a citation type has or has not been assigned. Citation classifications are based on the following two requirements: there needs to be a written indication that the statement supports or contrasts the cited paper; and there needs to be an indication that it provides evidence for this assertion.

Citation statement	Classification	Explanation
"In agreement with previous work (Nicholson et al., 2015), the trisomic clones showed similar aberrations, albeit to a lesser extent (Supplemental Figure S2B)."	Supporting	"In agreement with previous work" indicates support, while "the trisomic clones showed similar aberrations, albeit to a lesser degree (Supplemental Figure S2B)" provides evidence for this supporting statement.
"In contrast to several studies in anxious adults that examined amygdala activation to angry faces when awareness was not restricted (Phan, Fitzgerald, Nathan, & Tancer, 2006; Stein, Goldin, Sareen, Zorrilla, & Brown, 2002; Stein, Simmons, Feinstein, & Paulus, 2007), we found no group differences in amygdala activation."	Contrasting	"In contrast to several studies" indicates a contrast between the study and studies cited, while "we found no group differences in amygdala activation" indicates a difference in findings.
"The amygdala is a key structure within a complex circuit devoted to emotional interpretation, evaluation and response (Stein et al., 2002; Phan et al., 2006)."	Mentioning	This citation statement refers to Phan et al. (2006) without providing evidence that supports or contrasts the claims made in the cited study.
"In social cognition, the amygdala plays a central role in social reward anticipation and processing of ambiguity [87]. Consistent with these findings, amygdala involvement has been outlined as central in the pathophysiology of social anxiety disorders [27], [88]."	Mentioning	Here, the statement "consistent with these findings" sounds supportive, but, in fact, cites two previous studies: [87] and [27] without providing evidence for either. Such cites can be valuable, as they establish connections between observations made by others, but they do not provide primary evidence to support or contrast the cited studies. Hence, this citation statement is classified as mentioning.
"For example, a now-discredited article purporting a link between vaccination and autism (Wakefield et al., 1998) helped to dissuade many parents from obtaining vaccination for their children."	Mentioning	This citation statement describes the cited paper critically and with negative sentiment but there is no indication that it presents primary contrasting evidence, thus this statement is classified as mentioning.

immutable holdout set to avoid as much bias as possible. An n -fold cross-evaluation on the working set, for instance, would have been misleading because the distribution of the classes in this set was artificially modified to boost the classification accuracy of the less frequent classes.

Before reconciliation, the observed average interannotator agreement percentage was 78.5% in the open domain and close to 90% for batches in biomedicine. It is unclear what accounts for the difference. Reconciliation, further completed with expert review by core team members, resulted in highly consensual classification decisions, which contrast with typical multiround disagreement rates observed with sentiment classification. Athar (2014), for instance, reports Cohen's k annotator agreement of 0.675 and Ciancarini, Di Iorio et al. (2014) report $k = 0.13$ and $k = 0.15$ for the property groups covering *confirm/supports* and *critiques* citation classification labels. A custom open source document annotation web application, docanno (Nakayama, Kubo et al., 2018) was deployed to support the first round of annotations.

Overall, the creation of our current training and evaluation holdout data sets has been a major 2-year effort involving up to eight expert annotators and nearly 50,000 classified citation records. In addition to the class, each record includes the citation sentence, the full "snippet" (citation sentence plus previous and next sentences), the source and target DOI, the reference callout string, and the hierarchical list of section titles where the citation occurs.

2.6. Machine Learning Classifiers

Although deep learning text classifiers show very strong and stable results on imbalanced classification tasks compared with linear classifiers (Nizzoli, Avvenuti et al., 2019), our first experiments with an early training data set based on PLOS articles resulted in F -scores of 96.3% for mentioning citations, 55.3% for supporting, and 20.5% for contrasting. The initial accuracy for contrasting in particular raised concerns about the feasibility of the task itself at scale. We focused on multiple approaches to increase over time the accuracy of classifier for the two less frequent classes:

- Improving the classification architecture: After initial experiments with RNN (Recursive Neural Network) architectures such as BidGRU (Bidirectional Gated Recurrent Unit, an architecture similar to the approach of Cohan et al. (2019) for citation intent classification), we obtained significant improvements with the more recently introduced ELMo (Embeddings from Language Models) dynamic embeddings (Peters, Neumann et al., 2018) and an ensemble approach. Although the first experiments with BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang et al., 2019), a breakthrough architecture for NLP, were disappointing, fine-tuning SciBERT (a science-pretrained base BERT model) (Beltagy, Lo, & Cohan, 2019) led to the best results and is the current production architecture of the platform.
- Using oversampling and class weighting techniques: It is known that the techniques developed to address imbalanced classification in traditional machine learning can be applied successfully to deep learning too (Johnson & Khoshgoftaar, 2019). We introduced in our system oversampling of less frequent classes, class weighting, and metaclassification with three binary classifiers. These techniques provide some improvements, but they rely on empirical parameters that must be re-evaluated as the training data changes.
- Extending the training data for less frequent classes: As mentioned previously, we use an active learning approach to select the likely less frequent citation classes based on the

Table 2. Progress on classification results over approximately 1 year, evaluated on a fixed holdout set of 9,708 examples. In parallel with these various iterations on the classification algorithms, the training data was raised from 30,665 (initial evaluation with BidGRU) to 38,925 examples (last evaluation with SciBERT) via an active learning approach.

Approach	F-score		
	Contrasting	Supporting	Mentioning
BidGRU	.206	.554	.964
BidGRU + metaclassifier	.260	.590	.964
BidGRU + ELMo	.405	.590	.969
BidGRU + ELMo + ensemble (10 classifiers)	.460	.605	.972
SciBERT	.590	.648	.973
Observed distribution	0.8%	6.5%	92.6%

scores of the existing classifiers. By focusing on edge cases over months of manual annotations, we observed significant improvements in performance for predicting contrasting and supporting cases.

Because deep learning today is mostly an empirical effort, the improvements using the above-described techniques were driven experimentally and iteratively until reaching a plateau. Table 2 presents the model evaluation after iterations of the classification system over time using our fixed holdout set. Table 3 presents the evaluation metrics for the current SciBERT model. Reported scores are averaged over 10 runs. The *F*-score for the classification of “contrasting” was notably improved from 20.1% to 58.97%. The precision for predicting “contrasting” citations in particular reaches 85.19%, a very reliable level for such a rare class.

Given the unique nature of scite, there are a number of additional considerations. First, scaling is a key requirement of scite, which addresses the full corpus of scientific literature. While providing good results, the prediction with the ELMo approach is 20 times slower than with SciBERT, making it less attractive for our platform. Second, we have experimented with using section titles to improve classifications—for example, one might expect to find supporting and contrasting statements more often in the Results section of a paper and mentioning statements in the Introduction. Counterintuitively, including section titles in our model had no impact on *F*-scores, although it did slightly improve precision. It is unclear why including section titles failed to improve *F*-scores. However, it might relate to the challenge of correctly identifying and normalizing section titles from documents. Third, segmenting scientific text into sentences presents unique challenges due to the prevalence of abbreviations, nomenclatures, and

Table 3. Accuracy of SciBERT classifier, currently deployed on the scite platform, evaluated on a holdout set of 9,708 examples.

	Precision	Recall	F-score
Contrasting	.852	.451	.590
Supporting	.741	.576	.648
Mentioning	.962	.984	.973

Note: When deploying classification models in production, we balance the precision/recall so that all the classes have a precision higher than 80%.

mathematical equations. Finally, we experimented with various context windows (i.e., the amount of text used in the classification of a citation) but were only able to improve the F -score for the contrasting category by eight points by manually selecting the most relevant phrases in the context window. Automating this process might improve classifications, but doing so presents a significant technical challenge. Other possible improvements of the classifier include multitask training, refinement of classes, increase of training data via improved active learning techniques, and integration of categorical features in the transformer classifier architecture.

We believe that the specificity of our evidence-based citation classes, the size and the focus on the quality of our manually annotated data set (multiple rounds of blind annotations with final collective reconciliation), the customization and continuous improvement of a state of the art deep learning classifier, and finally the scale of our citation analysis distinguishes our work from existing developments in automatic citation analysis.

2.7. Citation Statement and Classification Pipeline

TEI XML data is parsed in Python using the BeautifulSoup library and further segmented into sentences using a combination of spaCy (Honnibal, Montani et al., 2018) and Natural Language Toolkit's Punkt Sentence Tokenizer (Bird, Klein, & Loper, 2009). These sentence segmentation candidates are then postprocessed with custom rules to better fit scientific texts, existing text structures, and inline markups. For instance, a sentence split is forbidden inside a reference callout, around common abbreviations not supported by the general-purpose sentence segmenters, or if it is conflicting with a list item, paragraph, or section break.

The implementation of the classifier is realized by a component we have named *Veracity*, which provides a custom set of deep learning classifiers built on top of the open source DeLFT library (Lopez, 2020d). *Veracity* is written in Python and employs Keras and TensorFlow for text classification. It runs on a single server with an NVIDIA GP102 (GeForce GTX 1080 Ti) graphics card with 3,584 CUDA cores. This single machine is capable of classifying all citation statements as they are processed. *Veracity* retrieves batches of text from the scite database that have yet to be classified, processes them, and updates the database with the results. When deploying classification models in production, we balance the precision/recall so that all the classes have a precision higher than 80%. For this purpose, we use the holdout data set to adjust the class weights at the prediction level. After evaluation, we can exploit all available labeled data to maximize the quality, and the holdout set captures a real-world distribution adapted to this final tuning.

2.8. User Interface

The resulting classified citations are stored and made available on the scite platform. Data from scite can be accessed in a number of ways (downloads of citations to a particular paper; the scite API, etc.). However, users will most commonly access scite through its web interface. Scite provides a number of core features, detailed below.

The scite report page (Figure 1) displays summary information about a given paper. All citations in the scite database to the paper are displayed, and users can filter results by classification (supporting, mentioning, contrasting), paper section (e.g., Introduction, Results), and the type of citing article (e.g., preprint, book, etc.). Users can also search for text within citation statements and surrounding citation context. For example, if a user wishes to examine how an article has been cited with respect to a given concept (e.g., fear), they can search for citation contexts that contain that key term. Each citation statement is accompanied by a classification

label, as well as an indication of how confident the model is of said classification. For example, a citation statement may be classified as supporting with 90% confidence, meaning that the model is 90% certain that the statement supports the target citation. Finally, each citation statement can be flagged by individual users as incorrect, so that users can report a classification as incorrect, as well as justify their objection. After a citation statement has been flagged as incorrect, it will be reviewed and verified by two independent reviewers, and, if both agree, the recommended change will be implemented. In this way, scite supplements machine learning with human interventions to ensure that citations are accurately classified. This is an important feature of scite that allows researchers to interact with the automated citation types, correcting classifications that might otherwise be difficult for a machine to classify. It also opens the possibility for authors and readers to add more nuance to citation typing by allowing them to annotate snippets.

To improve the utility and usability of the smart citation data, scite offers a wide variety of tools common to other citation platforms, such as Scopus and Web of Science and other information retrieval software. These include literature searching functionality for researchers to find supported and contrasted research, visualizations to see research in context, reference checking for automatically evaluating references with scite's data on an uploaded manuscript and more. Scite also offers plugins for popular web browsers and reference management software (e.g., Zotero) that allow easy access to scite reports and data in native research environments.

3. DISCUSSION

3.1. Research Applications

A number of researchers have already made use of scite for quantitative assessments of the literature. For example, Bordignon (2020) examined self-correction in the scientific record and operationalized “negative” citations as those that scite classified as contrasting. They found that negative citations are rare, even among works that have been retracted. In another example from our own group, Nicholson et al. (2020) examined scientific papers cited in Wikipedia articles and found that—like the scientific literature as a whole—the vast majority presented findings that have not been subsequently verified. Similar analyses could also be applied to articles in the popular press.

One can imagine a number of additional metascientific applications. For example, network analyses with directed graphs, valenced edges (by type of citation—supporting, contrasting, and mentioning), and individual papers as nodes could aid in understanding how various fields and subfields are related. A simplified form of this analysis is already implemented on the scite website (see Figure 3), but more complicated analyses that assess traditional network indices, such as centrality and clustering, could be easily implemented using standard software libraries and exports of data using the scite API.

3.2. Implications for Scholarly Publishers

There are a number of implications for scholarly publishers. At a very basic level, this is evident in the features that scite provides that are of particular use to publishers. For example, the scite Reference Check parses the reference list of an uploaded document and produces a report indicating how items in the list have been cited, flagging those that have been retracted or have otherwise been the subject of editorial concern. This type of screening can help publishers and editors ensure that articles appearing in their journals do not inadvertently cite

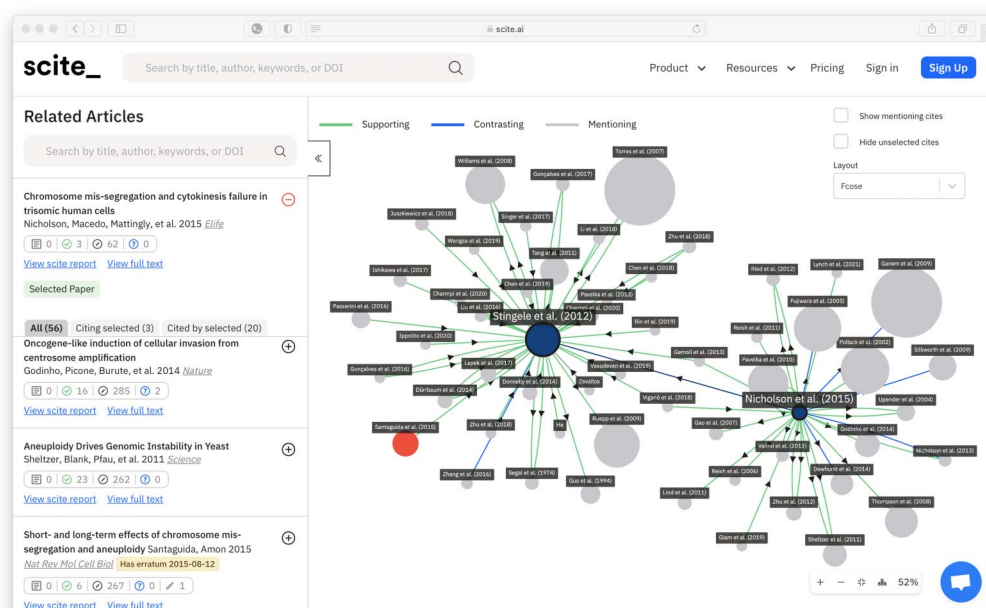


Figure 3. A citation network representation using the scite Visualization tool. The nodes represent individual papers, with the edges representing supporting (green) or contrasting (blue) citation statements. The graph is interactive and can be expanded and modified for other layouts. The interactive visualization can be accessed at the following link: <https://scite.ai/visualizations/global-analysis-of-genome-transcriptome-9L4dJr?dois%5B0%5D=10.1038%2Fmsb.2012.40&dois%5B1%5D=10.7554%2Felif.05068&focusedElement=10.7554%2Felif.05068>.

discredited works. Evidence in scite's own database indicates that this would solve a seemingly significant problem, as in 2019 alone nearly 6,000 published papers cited works that had been retracted prior to 2019. Given that over 95% of citations made to retracted articles are in error (Schneider, Ye et al., 2020), had the Reference Check tool been applied to these papers during the review process, the majority of these mistakes could have been caught.

However, there are additional implications for scholarly publishing that go beyond the features provided by scite. We believe that by providing insights into how articles are cited—rather than simply noting that the citation has occurred—scite can alter the way in which journals, institutions, and publishers are assessed. Scite provides journals and institutions with dashboards that indicate the extent to which papers with which they are associated have been supported or contrasted by subsequent research (Figure 4). Even without reliance on specific metrics, the approach that scite provides prompts the question: What if we normalized the assessment of journals, institutions and researchers in terms of how they were cited rather than the simple fact that they were cited alone?

3.3. Implications for Researchers

Given the fact that nearly 3 million scientific papers are published every year (Ware & Mabe, 2015), researchers increasingly report feeling overwhelmed by the amount of literature they must sift through as part of their regular workflow (Landhuis, 2016). Scite can help by assisting researchers in identifying relevant, reliable work that is narrowly tailored to their interests, as well as better understanding how a given paper fits into the broader context of the scientific literature. For example, one common technique for orienting oneself to new literature is to seek out the most highly cited papers in that area. If the context of those citations is also visible, the

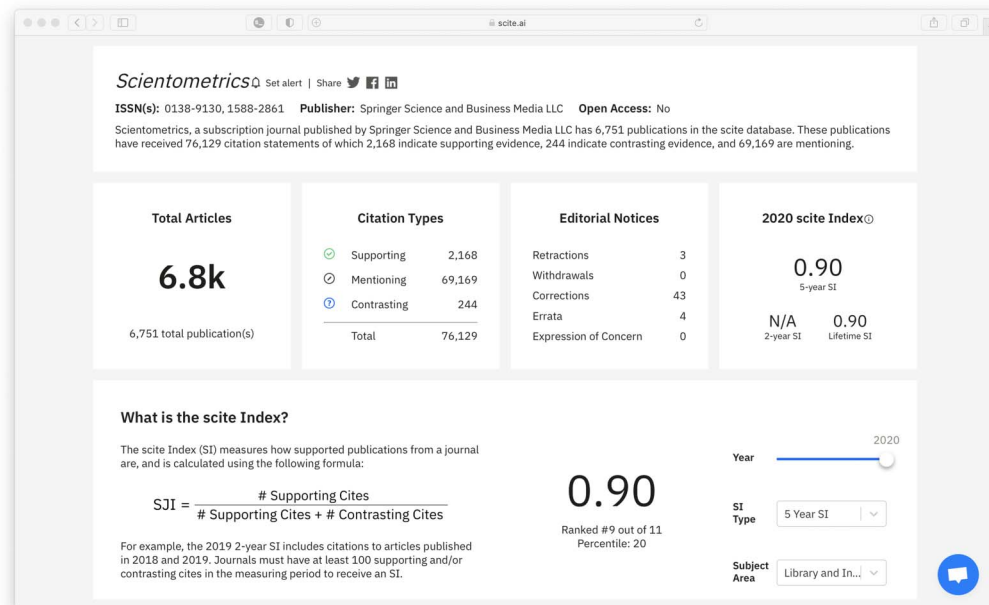


Figure 4. A scite Journal Dashboard showing the aggregate citation information at the journal level, including editorial notices and the scite Index, a journal metric that shows the ratio of supporting citations over supporting plus contrasting citations. Access to the journal dashboard in the figure and other journal dashboards is available here: <https://scite.ai/journals/0138-9130>.

value of a given paper can be more completely assessed and understood. There are, however, additional—although perhaps less obvious—implications. If citation types are easily visible, it is possible that researchers will be incentivized to make replication attempts easier (for example, by providing more explicit descriptions of methods or instruments) in the hope that their work will be replicated.

3.4. Limitations

At present, the biggest limitation for researchers using scite is the size of the database. At the time of this writing, scite has ingested over 880 million separate citation statements from over 25 million scholarly publications. However, there are over 70 million scientific publications in existence (Ware & Mabe, 2015); scite is constantly ingesting new papers from established sources and signing new licensing agreements with publishers, so this limitation should abate over time. However, given that the ingestion pipeline fails to identify approximately 30% of citation statements/references in PDF files (~5% in XML), the platform will necessarily contain fewer references than services such as Google Scholar and Web of Science, which do not rely on ingesting the full text of papers. Even if references are reliably extracted and matched with a DOI or directly provided by publishers, a reference is currently only visible on the scite platform if it is matched with at least one citation context in the body of the article. As such, the data provided by scite will necessarily miss a measurable percentage of citations to a given paper. We are working to address these limitations in two ways: First, we are working toward ingesting more full-text XML and improving our ability to detect document structure in PDFs. Second, we have recently supplemented our Smart Citation data with “traditional” citation metadata provided by Crossref (see “Without Citation Statements” shown in Figure 1), which surfaces references that we would otherwise miss. Indeed, this Crossref data now includes references from publishers with previously closed references such as Elsevier and the American

Chemical Society. These traditional citations can later be augmented to include citation contexts as we gain access to full text.

Another limitation is related to the classification of citations. First, as noted previously, the Veracity software does not perfectly classify citations. This can partly be explained by the fact that language in the (biomedical) sciences is little standardized (unlike law, where *shepardizing* is a standing term describing the “process of using a citator to discover the history of a case or statute to determine whether it is still good law”; see Lehman & Phelps, 2005). However, the accuracy of the classifier will likely increase over time as technology improves and the training data set increases in size. Second, the ontology currently employed by scite (supporting, mentioning, and contrasting) necessarily misses some nuance regarding how references are cited in scientific papers. One key example relates to what “counts” as a contrasting citation: At present, this category is limited to instances where new evidence is presented (e.g., a failed replication attempt or a difference in findings). However, it might also be appropriate to include conceptual and logical arguments against a given paper in this category. Moreover, in our system, the evidence behind the supporting or contrasting citation statements is not being assessed; thus a supporting citation statement might come from a paper where the experimental evidence is weak and vice versa. We do display the citation tallies that papers have received so that users can assess this but it would be exceedingly difficult to also classify the sample size, statistics, and other parameters that define how robust a finding is.

4. CONCLUSIONS

The automated extraction and analysis of scientific citations is a technically challenging task, but one whose time has come. By surfacing the context of citations rather than relying on their mere existence as an indication of a paper’s importance and impact, scite provides a novel approach to addressing pressing questions for the scientific community, including incentivizing replicable works, assessing an increasingly large body of literature, and quantitatively studying entire scientific fields.

ACKNOWLEDGMENTS

We would like to thank Yuri Lazebnik for his help in conceptualizing and building scite.

FUNDING INFORMATION

This work was supported by NIDA grant 4R44DA050155-02.

AUTHOR CONTRIBUTIONS

Josh M. Nicholson: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft, Writing—Review and editing. Milo Mordaunt: Data acquisition, Analysis and interpretation of data. Patrice Lopez: Conceptualization, Analysis and interpretation of data, Writing—original draft, Writing—Review and editing. Ashish Uppala: Analysis and interpretation of data, Writing—original draft, Writing—Review and editing. Domenic Rosati: Analysis and interpretation of data, Writing—original draft, Writing—Review and editing. Neves P. Rodrigues: Conceptualization. Sean C. Rife: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft, Writing—Review and editing. Peter Grabitz: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft, Writing—Review and editing.

COMPETING INTERESTS

The authors are shareholders and/or consultants or employees of Scite Inc.

DATA AVAILABILITY

Code used in the ingestion of manuscripts is available at <https://github.com/kermitt2/grobid>, <https://github.com/kermitt2/biblio-glutton>, and <https://github.com/kermitt2/Pub2TEI>. The classification of citation statements is performed by a modified version of DeLFT (<https://github.com/kermitt2/delft>). The training data used by the scite classifier is proprietary and not publicly available. The 880+ million citation statements are available at scite.ai but cannot be shared in full due to licensing arrangements made with publishers.

REFERENCES

- Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. *Proceedings of the ACL 2011 Student Session*, 81–87. Retrieved from <https://www.aclweb.org/anthology/P11-3015>
- Athar, A. (2014). *Sentiment analysis of scientific citations*. Technical Report (UCAM-CL-TR-856), University of Cambridge, Computer Laboratory. Retrieved from <https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-856.pdf>
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *ArXiv:1903.10676 [Cs]*. <https://arxiv.org/abs/1903.10676>. <https://doi.org/10.18653/v1/D19-1371>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1st ed.). O'Reilly.
- Bordignon, F. (2020). Self-correction of science: A comparative study of negative citations and post-publication peer review. *Scientometrics*, 124(2), 1225–1239. <https://doi.org/10.1007/s11192-020-03536-z>
- Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2014). Evaluating citation functions in CiTO: Cognitive issues. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, & A. Tordai (Eds.), *The Semantic Web: Trends and Challenges* (Vol. 8465, pp. 580–594). Springer International Publishing. https://doi.org/10.1007/978-3-319-07443-6_39
- Cohan, A., Ammar, W., van Zuilen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/N19-1361>
- Constantin, A. (2014). *Automatic structure and keyphrase analysis of scientific publications*. University of Manchester. https://www.research.manchester.ac.uk/portal/files/54553913/FULL_TEXT.PDF
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/N19-1423>
- Else, H. (2018). How Unpaywall is transforming open science. *Nature*, 560(7718), 290–291. <https://doi.org/10.1038/d41586-018-05968-3>, PubMed: 30111793
- Fujiwara, T., & Yamamoto, Y. (2015). Colil: A database and search service for citation contexts in the life sciences domain. *Journal of Biomedical Semantics*, 6(1), 38. <https://doi.org/10.1186/s13326-015-0037-x>, PubMed: 26500753
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111. <https://doi.org/10.1126/science.122.3159.108>, PubMed: 14385826
- Garfield, E. (1959). Letter to Dr. Joshua Lederberg, Stanford University. Retrieved from <https://www.garfield.library.upenn.edu/lederberg/052159.html>
- Garfield, E. (1964). Can Citation Indexing be Automated? Reprinted from M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation, symposium proceedings, Washington 1964* (pp. 189–192). National Bureau of Standards. Retrieved from <https://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf>
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178(4060), 471–479. <https://doi.org/10.1126/science.178.4060.471>, PubMed: 5079701
- Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, 339, b2680. <https://doi.org/10.1136/bmj.b2680>, PubMed: 19622839
- Halevi, G., & Schimming, L. (2018). An initiative to track sentiments in altmetrics. *Journal of Altmetrics*, 1(1), 2. <https://doi.org/10.29024/joa.1>
- 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>
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572. <https://doi.org/10.1073/pnas.0507655102>, PubMed: 16275915
- Honnibal, M., Montani, I., Honnibal, M., Peters, H., Samsonov, M., ... Patel, A. (2018). *Explosion/paCy: V2.0.11: Alpha Vietnamese support, fixes to vectors, improved errors and more*. Zenodo. <https://doi.org/10.5281/ZENODO.1212304>
- Initiative for Open Citations. (2017). <https://i4oc.org/>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27. <https://doi.org/10.1186/s40537-019-0192-5>
- Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535(7612), 457–458. <https://doi.org/10.1038/nj7612-457a>, PubMed: 27453968
- Lehman, J., & Phelps, S. (2005). Shepardizing. In *West's encyclopedia of American law* (2nd ed., vol. 9, p. 162). Detroit: Thomson/Gale.
- Letrud, K., & Hernes, S. (2019). Affirmative citation bias in scientific myth debunking: A three-in-one case study. *PLOS ONE*, 14(9), e0222213. <https://doi.org/10.1371/journal.pone.0222213>, PubMed: 31498834
- Leung, P. T. M., Macdonald, E. M., Stanbrook, M. B., Dhalla, I. A., & Juurlink, D. N. (2017). A 1980 letter on the risk of opioid addiction. *New England Journal of Medicine*, 376(22), 2194–2195. <https://doi.org/10.1056/NEJMc1700150>
- Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *ArXiv:1911.02782 [Cs]*. <https://arxiv.org/abs/1911.02782>
- Lopez, P. (2020a). *GROBID* [source code]. Retrieved from <https://github.com/kermitt2/grobid>

- Lopez, P. (2020b). *Pub2TEI* [source code]. Retrieved from <https://github.com/kermitt2/Pub2TEI>
- Lopez, P. (2020c). *biblio-glutton* [source code]. Retrieved from <https://github.com/kermitt2/biblio-glutton>
- Lopez, P. (2020d). *delft* [source code]. Retrieved from <https://github.com/kermitt2/delft>
- Murray, D., Lamers, W., Boyack, K., Larivière, V., & Sugimoto, C. R. (2019). Measuring disagreement in science. *17th International Conference on Scientometrics & Informetrics* (pp. 2370–2375). September 2–5, 2019. Retrieved from <https://crctcs.openum.ca/files/sites/60/2019/10/ISSI2019-measuring-disagreement-in-science.pdf>
- Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text annotation tool for humans. <https://github.com/doccano/doccano>
- Nicholson, J. M., Macedo, J. C., Mattingly, A. J., Wangsa, D., Camps, J., ... Cimini, D. (2015). Chromosome mis-segregation and cytokinesis failure in trisomic human cells. *eLife*, 4, e05068. <https://doi.org/10.7554/eLife.05068>, PubMed: 25942454
- Nicholson, J. M., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M., & Rife, S. C. (2020). Measuring the quality of scientific references in Wikipedia: An analysis of more than 115M citations to over 800 000 scientific articles. *FEBS Journal*, 288(14), 4242–4248. <https://doi.org/10.1111/febs.15608>
- Nizzoli, L., Avvenuti, M., Cresci, S., & Tesconi, M. (2019). Extremist propaganda tweet classification with deep learning in realistic scenarios. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19* (pp. 203–204). <https://doi.org/10.1145/3292522.3326050>
- Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17, 33–43. <https://doi.org/10.1016/j.websem.2012.08.001>
- Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., ... Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1202>
- Phan, K. L., Fitzgerald, D. A., Nathan, P. J., & Tancer, M. E. (2006). Association between amygdala hyperactivity to harsh faces and severity of social anxiety in generalized social phobia. *Biological Psychiatry*, 59(5), 424–429. <https://doi.org/10.1016/j.biopsych.2005.08.012>, PubMed: 16256956
- Piowar, H., Priem, J., & Orr, R. (2019). The future of OA: A large-scale analysis projecting Open Access publication and readership [Preprint]. *Scientific Communication and Education*. <https://doi.org/10.1101/795310>
- PLOS. (2015). *Rich_citations* [source code]. Retrieved from https://github.com/PLOS/rich_citations
- Porter, J., & Jick, H. (1980). Addiction rare in patients treated with narcotics. *New England Journal of Medicine*, 302(2), 123. <https://doi.org/10.1056/NEJM198001103020221>
- Pride, D., Knoth, P., & Harag, J. (2019). ACT: An annotation platform for citation typing at scale. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 329–330). <https://doi.org/10.1109/JCDL.2019.00055>
- Schneider, J., Ye, D., Hill, A. M., & Whitehorn, A. S. (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, 125(3), 2877–2913. <https://doi.org/10.1007/s11192-020-03631-1>
- Stein, M. B., Goldin, P. R., Sareen, J., Zorrilla, L. T., & Brown, G. G. (2002). Increased amygdala activation to angry and contemptuous faces in generalized social phobia. *Archives of General Psychiatry*, 59, 1027–1034. <https://doi.org/10.1001/archpsyc.59.11.1027>, PubMed: 12418936
- Stein, M. B., Simmons, A. N., Feinstein, J. S., & Paulus, M. P. (2007). Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. *American Journal of Psychiatry*, 164(2), 318–327. <https://doi.org/10.1176/ajp.2007.164.2.318>
- Suelzer, E. M., Deal, J., Hanus, K. L., Ruggeri, B., Sieracki, R., & Witkowski, E. (2019). Assessment of citations of the retracted article by Wakefield et al with fraudulent claims of an association between vaccination and autism. *JAMA Network Open*, 2(11), e1915552. <https://doi.org/10.1001/jamanetworkopen.2019.15552>, PubMed: 31730183
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). <https://doi.org/10.3115/1610075.1610091>
- Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. *ArXiv:1802.01168 [CS]*. <https://arxiv.org/abs/1802.01168>. <https://doi.org/10.1145/3197026.3197048>
- Viganó, C., von Schubert, C., Ahrné, E., Schmidt, A., Lorber, T., ... Nigg, E. A. (2018). Quantitative proteomic and phosphoproteomic comparison of human colon cancer DLD-1 cells differing in ploidy and chromosome stability. *Molecular Biology of the Cell*, 29(9), 1031–1047. <https://doi.org/10.1091/mbc.E17-10-0577>, PubMed: 29496963
- Volanakis, A., & Krawczyk, K. (2018). SciRide Finder: A citation-based paradigm in biomedical literature search. *Scientific Reports*, 8(1), 6193. <https://doi.org/10.1038/s41598-018-24571-0>, PubMed: 29670147
- Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., ... Walker-Smith, J. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103), 637–641. [https://doi.org/10.1016/S0140-6736\(97\)11096-0](https://doi.org/10.1016/S0140-6736(97)11096-0), PubMed: 9500320
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., ... Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *ArXiv:2004.10706 [CS]*. <https://arxiv.org/abs/2004.10706>
- Ware, M., & Mabe, M. (2015). *The STM Report: An overview of scientific and scholarly journal publishing*, p. 181. The Hague: International Association of Scientific, Technical and Medical Publishers.
- Yan, E., Chen, Z., & Li, K. (2020). The relationship between journal citation impact and citation sentiment: A study of 32 million citations in PubMed Central. *Quantitative Science Studies*, 1(2), 664–674. https://doi.org/10.1162/qss_a_00040
- Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019). A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*, 52(3), 1805–1838. <https://doi.org/10.1007/s10462-017-9597-8>