RESEARCH ARTICLE

A meta-analysis of semantic classification of citations

Suchetha N. Kunnath[1], Drahomira Herrmannova[2], David Pride[1], and Petr Knoth[1]

[1]Knowledge Media Institute (KMi), The Open University, Milton Keynes, UK
[2]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

Keywords: citation classification, citation context, citation function, citation importance, citation polarity, citation type

ABSTRACT

The aim of this literature review is to examine the current state of the art in the area of citation classification. In particular, we investigate the approaches for characterizing citations based on their semantic type. We conduct this literature review as a meta-analysis covering 60 scholarly articles in this domain. Although we included some of the manual pioneering works in this review, more emphasis is placed on the later automated methods, which use Machine Learning and Natural Language Processing (NLP) for analyzing the fine-grained linguistic features in the surrounding text of citations. The sections are organized based on the steps involved in the pipeline for citation classification. Specifically, we explore the existing classification schemes, data sets, preprocessing methods, extraction of contextual and noncontextual features, and the different types of classifiers and evaluation approaches. The review highlights the importance of identifying the citation types for research evaluation, the challenges faced by the researchers in the process, and the existing research gaps in this field.

1. INTRODUCTION

Citation analysis has been a subject of study for several decades, with the work of Garfield (1972) being among the most pioneering. One of the primary motivations for studies related to bibliographic references is to identify methods for research assessment and evaluation (Swales, 1986). Existing methods using citation impact indicators such as the *h*-index and Journal Impact Factors (JIFs), which are based on citation frequency, have been used alongside the earlier peer-reviewing approaches for research evaluation (Aksnes, Langfeldt, & Wouters, 2019). Traditional use of citation counts alone as an indicator for measuring the scientific impact of research publications, researchers, and research institutions has been widely criticized in the past (Kaplan, 1965; Moravcsik & Murugesan, 1975). The San Francisco Declaration on Research Assessment (DORA)[1] released in 2013 includes 18 recommendations for improving research evaluation methods to mitigate the limitations of the citation count based impact assessment methods. According to Garfield (1972), "... citation frequency is, of course, a function of many variables besides scientific merit ...." Some of these factors that affect citation frequency are time since publication, field, journal, article, author or reader, and the

[1] https://sfdora.org/read/

an open access journal

Citation: Kunnath, S. N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, *2*(4), 1170–1215. https://doi.org/10.1162/qss_a_00159

DOI: https://doi.org/10.1162/qss_a_00159

Peer Review: https://publons.com/publon/10.1162/qss_a_00159

Received: 19 February 2021
Accepted: 10 September 2021

Corresponding Author:
Suchetha N. Kunnath
snk56@open.ac.uk

Handling Editor:
Ludo Waltman

The MIT Press

publication's availability (Bornmann & Daniel, 2008). How to weigh such individual factors is still unclear when using citation measures for evaluating research (Garfield, 1979).

Earlier methods based on citation counting for assessing the scientific impact of publications treat all citations with equal weights, regardless of their function. A number of researchers have argued that this oversimplification is detrimental to the use of citation data in research evaluation systems (Jha, Jbara et al., 2017; Jurgens, Kumar et al., 2018; Zhu, Turney et al., 2015). For instance, a citation that criticizes a work has a different influence than a citation used as a starting point for new research (Hernández-Álvarez, Gomez Soriano, & Martínez-Barco, 2017). Abu-Jbara, Ezra, and Radev (2013) state that the number of citations received is just an indication of the productivity of a researcher and the publicity the work received; it does not convey any information about the quality of the research itself. Besides, overview papers often generate greater citation counts than some of the seminal publications (Herrmannova, Patton et al., 2018; Ioannidis, 2006). Negative citations, self-citations, and citations to methodological papers all raise questions regarding the validity of using citation counts for research evaluation (Garfield, 1979). More recent publications that make independent scientific contributions may not have yet received enough citations to be considered as impactful (Herrmannova et al., 2018). Additionally, Gilbert (1977) argues that, instead of a research evaluation purpose, citations act as a tool for persuasion, convincing the readers about the validity and significance of the presented claims. This illustrates the potential of these tools in improving bibliometric research evaluation methods such that the citation type is also taken into account.

The apprehension concerning the appropriateness and the reliability of methodologies involving mere citation counting in the context of research evaluation constitutes a key application area that encouraged the development of techniques for identifying the functional typology of citations. A pioneering work by Moravcsik and Murugesan (1975) found that out of 575 bibliographic references from 30 articles, 40% of citations were perfunctory and 33% of them were redundant, raising concerns about using citation counts as a quality measure. Research in this direction is often motivated by the observation that readers interested in not just how many times a work is cited but also why it is being cited (Lauscher, Glavaš et al., 2017). However, Nakov, Schwartz et al. (2004) show that there are a variety of other application areas, including document summarization, document indexing and retrieval and monitoring research trends, that can be seen as beneficiaries of citation classification technology.

In this meta-analysis, we review existing research on semantic classification of citations. Specifically, we focus on studies that exploit citation context (i.e., the textual fragment surrounding a citation marker within the cited paper) to determine the citation type. Unlike the previous survey papers in this domain (Bornmann & Daniel, 2008; Hernández-Álvarez & Gomez, 2016; Tahamtan & Bornmann, 2019), we focused not just on the available methods for citation classification and the citation context analysis but also the different phases of the general pipeline for the task. The existing papers are systematically reviewed based on the steps involved in citation classification. More emphasis is placed on the later automated methods than on the earlier manual work for citation classification.

This paper is organized as follows: Section 2 describes the process of citation classification, important terminologies, applications, and challenges in this area. Section 3 explains the methods we used for collecting research papers for this meta-analysis. Sections 4 and 5 review the popular classification schemes and the data sets. This is followed by examining methods used for the different steps involved in the automatic citation classification, namely preprocessing, important feature identification, classification, and evaluation. Section 10 describes the open competitions in this domain.

2. CITATION CLASSIFICATION

Research publications are not standalone entities, but rather individual pieces of literature pointing to prior research. This connection between the research publications is accomplished through the use of citations, which act as a bridge between the citing and the cited document. The reason or motivation for citing a paper has been studied extensively by sociologists of science and information scientists in the past (Cano, 1989; Gilbert, 1977; Moravcsik & Murugesan, 1975; Oppenheim & Renn, 1978). Garfield (1965) in his pioneering work identifies 15 reasons for citing a paper, a few of which are "Paying homage to pioneers, Giving credit for related work, Identifying method, equipment etc., Providing background reading" and so forth. All these studies developed taxonomies for characterizing citations aimed at identifying the social functions that reference serves and determining how important it is to the citing author in order to give insight into authors' citing practices (Radoulov, 2008). Earlier methods used either surveys of published authors (Brooks, 1985; Cano, 1989) or the expertise of the analysts (Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975) to decode the implicit aspects of citations from the text surrounding the reference (Sula & Miller, 2014). However, little attention was given to analyzing the scientific content of the citation context.

The citation classification problem from a discourse analyst point of view was later studied by Swales (1986), Teufel, Siddharthan, and Tidhar (2006b), and White (2004). Here, the explicitly mentioned words or phrases surrounding the citation are analyzed to interpret the author's intentions for citing a document (White, 2004). To this end, several taxonomies, from the very generic to the more fine grained, were developed reflecting on citation types from a range of perspectives. These include understanding citation functions, which constitute the roles or purposes associated with a citation, by examining the citation context (Cohan, Ammar et al., 2019; Garzone & Mercer, 2000; Jurgens et al., 2018; Teufel et al., 2006b); citation polarity or sentiment, which gives insight into the author's disposition towards the cited document (Hernández-Álvarez et al., 2017; Lauscher et al., 2017); and citation importance, where the citations are grouped based on how influential/important they are to the cited document (Pride & Knoth, 2017b; Valenzuela, Ha, & Etzioni, 2015; Zhu et al., 2015).

Progress in research related to the fields of Machine Learning and NLP resulted in the development of automatic methods for evaluating citation context and extraction of textual and nontextual features, followed by the classification of citations. Figure 4 represents the general steps involved in citation classification. In this literature review, we intend to explore the literature that examines the qualitative aspects of citation classification; citation function and importance. This meta-analysis also covers previous research related to each of the steps indicated in Figure 4 and inspects the different techniques used by past studies. In the following section, we describe the terminologies associated with citation classification in the context of a discursive relationship between the cited and the citing text. This is followed by the subsections, challenges and applications of automatic citation classification methods.

2.1. Terminology

The following are the key terms associated with this meta-analysis:

- **Citing Sentence/Citance** represents the sentence in the citing paper which contain the citations.
- **Citation Context** constitutes the citing text as well as the related text surrounding the citation that the citing authors use to describe the cited paper.
- **Citation Context Analysis** facilitates the syntactic and semantic analysis of the contents of the citation context to understand how and why authors discuss others, research work.

- **Citation Classifier** predicts the function, polarity or importance of citations, given the citation context or the citing sentence. The function here represents the different aspects of citation, for instance, purpose, intent, or reason for citing. Polarity represents the author's sentiment towards the citation. Importance is a measure of how influential the cited research work is.
- **Citation Type** is any overarching term for any semantic type, including function, polarity, importance, intent etc.
- **Citation Classification Scheme** specifies the different categories (and their definition) used for classifying citations.

2.2. Challenges

Classifying citations based on their type is not a trivial task. First, the citing sentence might not always explicitly contain the necessary semantic cues enabling us to determine the citation type. Second, authors frequently refer to a previously cited document further on in their manuscript using named entities, such as names of the used methods, tools or data sets, without explicitly mentioning the citation (Kaplan, Tokunaga, & Teufel, 2016). Disregarding such implicit citations results in an information loss when characterizing citations (Athar & Teufel, 2012b). Occasionally, authors use exaggerated praise to hide criticism, thus avoiding negative citations, and show reluctance to acknowledge using a specific method from previous research (Teufel, Siddharthan, & Tidhar, 2006a). Developing a classification scheme that can successfully capture the broad range of citation functions too is challenging. Classification schemes often range from the rather abstract to the fine grained. Although the abstract taxonomies are too general to capture all the specific information (Radoulov, 2008), the interannotator agreement decreases substantially in the case of the fine-grained schemes, with the annotators experiencing difficulties in choosing between similar or overlapping categories (Agarwal, Choubey, & Yu, 2010; Hernández-Álvarez, Gómez et al., 2016; Teufel et al., 2006a). Occasionally, the granularity of the fine-grained schemes is reduced due to the complications associated with such annotation procedures (Fisas, Ronzano, & Saggion, 2016). Additionally, most of the existing data sets for citation classification are manually annotated by domain experts, which is hugely time consuming and therefore expensive, and also potentially subjective (Bakhti, Niu, & Nyamawe, 2018).

Progress in this field has been hampered by the lack of annotated corpora large enough to generalize the task, and irrespective of the domain (Hernández-Álvarez & Gomez, 2016; Hernández-Álvarez et al., 2016; Radoulov, 2008). Nonreuse of the existing data sets, annotation schemes and the use of different feature sets and different classifiers makes the accurate comparison of findings from the current state of the art a rather problematic task (Jochim & Schütze, 2012). Moreover, the lack of methods for the formal comparison and evaluation of the citation classification systems makes it difficult to gauge the advancement of the state of the art (Kunnath, Pride et al., 2020). The domain-specific nature of existing data sets means the application of such corpora across multiple disciplines is a rather difficult prospect (White, 2004). Besides, considerable dissimilarities in the corpus and classification schemes and the classifiers used for the experiments means reproducing earlier results using a new corpus is challenging. The data sets developed for citation classification are highly skewed, with the majority of the instances belonging to the category corresponding to the background work, perfunctory or neutral category (Dong & Schäfer, 2011; Fisas et al., 2016; Jurgens et al., 2018). Often supervised learning methods for citation classification fail to categorize citations to the minority classes, which are of more importance in this task (Dong & Schäfer, 2011).

2.3. Applications

The taxonomy used for classifying citations according to different categories varies depending on the application for which the system is utilized. Some of the important applications that make use of citation typing information are research evaluation frameworks, summary generation systems, citation indexers, and so forth. Tools for analyzing citation purposes can help the funding agencies' decisions for ranking research papers, researchers, and Universities (Abu-Jbara et al., 2013). According to Xu et al. (2013), "... typed citations help identify seminal work and the main research paradigms of a field ...". Athar and Teufel (2012a) propose using citation sentiment to understand the research gaps and issues with the existing approaches. Valenzuela et al. (2015) incorporate the citation importance classification information to a scientific literature search engine for identifying the most important papers for a given cited work. In most cases, the detection of citation type is a prerequisite for many applications concerning scholarly publications (Radoulov, 2008). For instance, Nanba et al. (2000) classify the citation types for automatically generating review articles.

To extract the most representative subset for citation-based summary generation, Abu-Jbara and Radev (2011) classify the initial filtered citing sentences based on the five function types: *Background*, *Problem Statement*, *Method*, *Results*, and *Limitations*. Fisas et al. (2016) introduced a multilayer corpus with annotations for citation purpose as well as sentence relevance for scientific document summary. The extraction of hedging cues for detecting the fine-grained citation types was explored by Di Marco et al. (2006) to develop citation indexing tool for biomedical articles. Le et al. (2006) propose methods for integrating citation type detection as an initial step for discovering emerging trends. Schäfer and Kasterka (2010) developed a citation graph visualization tool based on typed citations to aid literature reviewing. Scite_[2], a commercial online platform, which does not have their training data and models openly available, identifies how citations are cited in research papers using the citation context for information retrieval. Table 1 shows the percentage distribution of papers and their corresponding applications out of the total number of papers reviewed for this meta-analysis. The values show that the majority of papers propose citation classification as a method for research evaluation.

3. SURVEY METHODOLOGY

In this meta-analysis, we review critical literature in the area of citation classification. The following reasons motivated us to do this literature review:

- Identify key papers of the field.
- Review trends, classification schemes, data sets and methods used by the existing systems.
- Comprehend the limitations and the research gaps.
- Determine the possible research directions in the domain.

The following subsection describes the method used for selecting the scientific publications for this survey.

3.1. Data Collection

Figure 1 illustrates the steps involved in the collection of research papers for this literature review. Initially, we identified the following keywords related to citation classification:

- Citation classification
- Citation function

[2] https://scite.ai/

Table 1. Reviewed papers with domain specific applications for citation classification

Application	Paper	%+
Information retrieval	Garzone and Mercer (2000)	11.6%
	Di Marco, Kroon, and Mercer (2006)	
	Schäfer and Kasterka (2010)	
	Agarwal et al. (2010)	
	Bertin and Atanassova (2012)	
	Xu, Martin, and Mahidadia (2013)	
	Valenzuela et al. (2015)	
Summarization	Nanba, Kando, and Okumura (2000)	6.6%
	Abu-Jbara and Radev (2011)	
	Fisas et al. (2016)	
	Jha et al. (2017)	
Research trend detection	Le, Ho, and Nakamori (2006)	8.3%
	Jha et al. (2017)	
	Hassan, Akram, and Haddawy (2017)	
	Hassan, Safder et al. (2018)	
	Jurgens et al. (2018)	
Research evaluation	Moravcsik and Murugesan (1975)	28.3%
	Chubin and Moitra (1975)	
	Spiegel-Rösing (1977)	
	Brooks (1985)	
	Cano (1989)	
	Abu-Jbara et al. (2013)	
	Valenzuela et al. (2015)	
	Zhu et al. (2015)	
	Hernández-Álvarez et al. (2017)	
	Lauscher et al. (2017)	
	Hassan et al. (2017)	
	Jurgens et al. (2018)	
	Cohan et al. (2019)	
	Qayyum and Afzal (2019)	
	Yousif, Niu et al. (2019)	

Table 1. (*continued*)

Application	Paper	%+
	Nazir, Asif, and Ahmad (2020a)	
	Pride and Knoth (2020)	
Venue evolution	Jurgens et al. (2018)	1.6%

+ Out of total papers reviewed.

- Citation polarity
- Citation sentiment
- Citation importance
- Citation context classification
- Citation motivation
- Citation intent
- Citation purpose
- Citation behavior and
- Citation annotation

Using these keywords, we queried the academic search engines Google Scholar[3], Scopus[4], ScienceDirect[5], CORE[6], and ACM Digital Library[7]. Additionally, we also searched for research papers using more generic terms such as "Citation Context Analysis" and "Citation Analysis." However, searching using these terms resulted in a far too broad set of research papers, beyond the scope of this literature review. For retrieving the relevant literature, we only selected papers from the top five pages from the above sources. In the final step, the collected papers were filtered by removing all the research publications, which are outside the scope of this meta-analysis. Moreover, we populated the list with papers from the reference sections of the initially collected papers that are significant and not already in the list.

Figure 2 presents the research papers included in this literature review for citation function and importance classification and the year in which these were published. The 60 papers represented in the diagram discuss taxonomies, data sets, or methods for citation classification. Nearly 87% of the documents reviewed are from post-2000, and we focused more on research corresponding to the automated approaches for citation classification. Additionally, we also review papers that discuss prerequisite steps such as scientific text extraction and preprocessing for citation classification. Table 2 shows the distribution of topics concerning the final list of papers cited in this survey paper. Nearly 42% of the papers discussed methods for citation function (purpose, polarity, or both). The reviewed documents for citation function and importance classification uses the following approaches: Manual, Rule-based, Machine Learning, and Deep Learning, the percentage distribution of which is represented in Figure 3.

[3] https://scholar.google.com/
[4] https://www.scopus.com/home.uri
[5] https://www.sciencedirect.com/
[6] https://core.ac.uk/
[7] https://dl.acm.org/

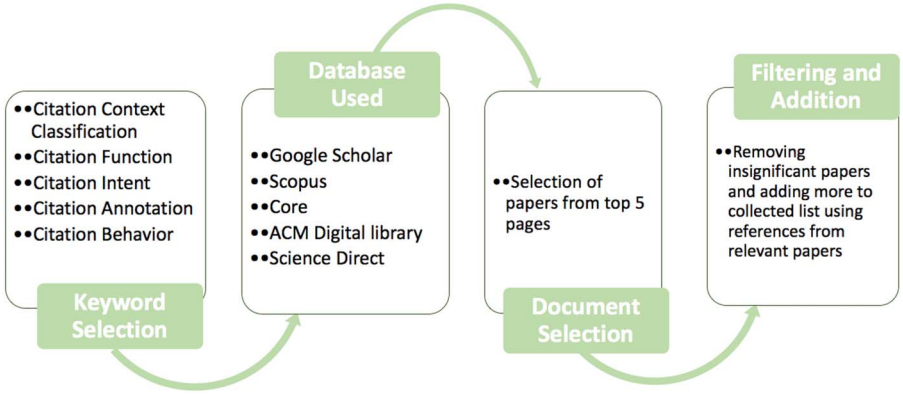

Figure 1. Steps involved in the collection of research papers for this survey.

4. CLASSIFICATION SCHEMES

This section describes the classification taxonomies associated with the existing systems for citation classification. In the first subsection, we will describe some of the early classification schemes for manual classification of the citations. This is followed by subsections on citation importance and citation function schemes, both of which are utilized by the recent automated approaches.

4.1. Early Research in Citation Classification

The earliest work in citation classification is attributed to Garfield (1965), who laid the foundation of this domain by proposing 15 reasons why authors cite a paper. However, Garfield just defined the different categories, and did not conduct in-depth research regarding the occurrence of different citation functions with respect to a paper. With the aim of determining

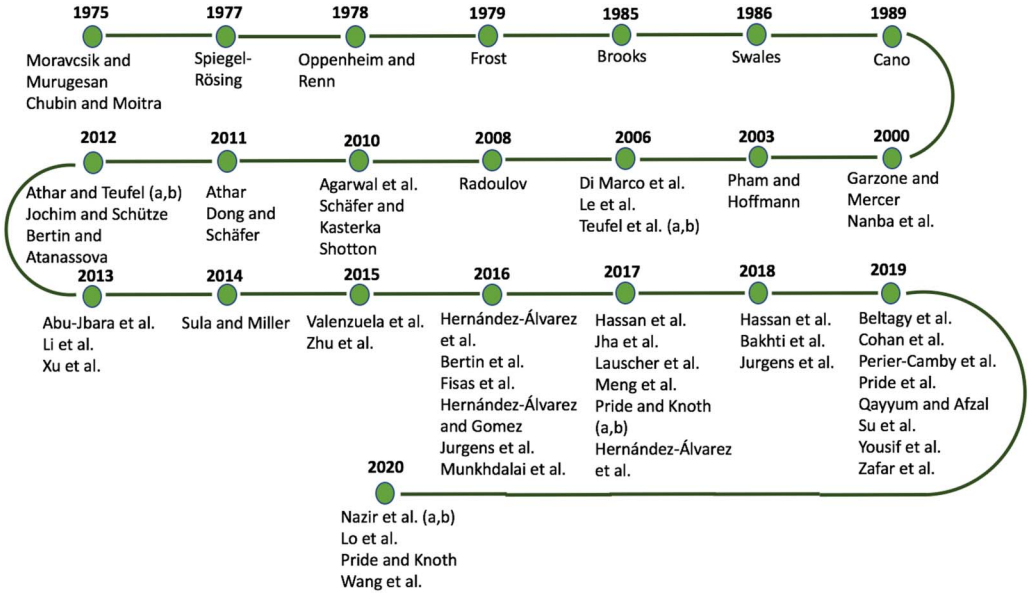

Figure 2. Timeline of the papers reviewed for this meta-analysis.

Table 2. Topical distribution of papers cited in this survey paper

Citation function & polarity	Citation importance	Citation analysis	Data set	Tools	Shared task	Others
41.7%	11.5%	9.4%	8.3%	7.3%	9.4%	12.5%

the citation type by analyzing the content text, Moravcsik and Murugesan (1975) developed a four-dimensional mutually exclusive annotation scheme using 30 articles from theoretical high-energy physics, the first of its kind, for classifying citations based on their quality and functions. Chubin and Moitra (1975) further extended this approach to address the limitations concerning the generalizability of Moravscik and Murugesan's scheme by introducing a hierarchical annotation schema featuring six basic classes. Using 66 articles from the journal *Science Studies*, Spiegel-Rösing (1977) introduced a classification scheme for research outside of Physics. Out of the 2,309 citations, 80% of them belonged to the category corresponding to cited source used for substantiating a statement or assumption. Frost (1979) addressed the question of finding classification functions common to both scientific and literary research. As subjective opinion has more importance than factual evidence in literary research, Frost (1979) designed a classification scheme specifically for humanities. Such interdisciplinary and intradisciplinary variations in citation functions have been observed by researchers (Chubin & Moitra, 1975; Harwood, 2009). Oppenheim and Renn (1978) studied 23 highly cited pre-1930 papers using 978 citing papers for identifying the authors' reasons for citing these articles. They used seven categories for classifying reasons for citation and came to the conclusion that nearly 40% of the highly cited articles are referenced for historical reasons.

Table 3 shows some of the initial schemes used for citation function classification. Earlier classification schemes suffered several downsides. For instance, the annotation scheme developed by Chubin and Moitra (1975) considered only one category for a reference, no matter in how many contexts the citation appeared in the paper. The limited availability of full text resulted in confining the research to specific journals and analysis of few references and articles. Also, the manual classification of citations to their respective functions requires reading the full text and annotations by subject experts (Hou, Li, & Niu, 2011). Moreover, most of the the distinction of citations resulting from the earlier taxonomies is sociologically oriented to a greater extent and is difficult to use for practical applications (Swales, 1986; Teufel et al., 2006a). None of the schemes mentioned here makes any differentiation between self-citations: a way to manipulate citation counts and citations to others' work (Swales, 1986). Swales (1986) raises the concern as to whether it is possible to determine the intent for citing by

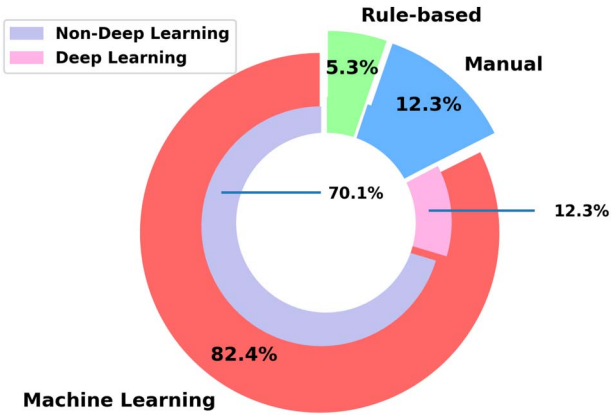

Figure 3. Distribution of citation classification methods used by the reviewed research papers.

Table 3. Early citation function annotation schemes

Authors	Classification scheme	Data source	Data size
Moravcsik and Murugesan (1975)	**Conceptual** or **Operational Use**	Theoretical high-energy physics published in *Physical Review* from 1968 to 1972 (inclusive)	30 articles
	Evolutionary or **Juxtapositional**		575 references
	Organic or **Perfunctory**		
	Confirmative or **Negational**		
Chubin and Moitra (1975)	**Affirmative**: (1) Basic, (2) Subsidiary, (3) Additional, (4) Perfunctory	33 research notes published in *Physical Review Letters* and *Physical Review B*	43 articles
	Negative: (1) Partial, (2) Total	10 full length articles from *Physics Review* and *Nuclear Physics* (January 1968–September 1969)	
Frost (1979)	**Primary Source**: (1) Supporting Factual Evidence, (2) Supporting Circumstantial Evidence	German Literature articles from journals *The Germanic Review*, *Euphorian*, and *Weimarer Beitrage* from years 1935, 1956, 1972	60 articles
	Secondary Source: (1) Acknowledging Pioneering works, (2) Indicating views on topic, (3) Refer to terms/symbols, (4) Support opinion, (5) Support facts, (6) Improvement of Idea, (7) Acknowledge Intellectual Indebtedness, (8) Disagree with opinion, (9) Disagree with facts, (10) Expressing Mixed Opinion		
	Either Primary or Secondary: (11) Refer to further reading, (12) Provide Bibliographic Information		

Table 3. (*continued*)

Authors	Classification scheme	Data source	Data size
Spiegel-Rösing (1977)	(1) Citation mentioned in Introduction/Discussion (2) Cited source is the specific point of departure for the research question (3) Cited source contains the concepts, definitions, interpretations used (4) Cited source contains data used by citing text (5) Cited source contains the data used for comparative purpose (6) Cited source contains data and material (from other disciplines than citing article) (7) Cited source contains method used (8) Cited source substantiates a statement or assumption (9) Cited source is positively evaluated (10) Cited source is negatively evaluated (11) Results of citing article prove,verify, substantiate data or interpretation of cited source (12) Results of citing article disprove, put into question the data as interpretation of cited source (13) Results of citing article furnish a new interpretation/explanation of data of cited source	Social Science Citation Index (1972–1975)	66 articles 2309 citations
Oppenheim and Renn (1978)	(1) Historical Background (2) Description of other relevant work (3) Supplying information or data, not for comparison (4) Supplying information or data, for comparison (5) Use of theoretical equation (6) Use of methodology (7) Theory or methods not applicable	Physics and Physical Chemistry	23 source articles 978 citing articles (1974–1975)
Brooks (1985)	(1) Currency Scale (2) Negative Credit (3) Operational Information (4) Persuasiveness (5) Positive Credit (6) Reader Alert (7) Social Consensus	Multidisciplinary	Papers by 26 faculties of University of Iowa

analyzing the citation context, as "... the reason why an author cites as he does must remain a matter for conjecture ...." A study by Cano (1989) on Moravscik and Muregesan's scheme shows that the annotation of citations by authors themselves to multiple classes was paired within the expected dichotomous categories. According to the author, Moravscik and Murugesan's citation behavior model could not fit in the "... research subject's perception of their use of information ...."

4.2. Citation Importance

Earlier research on citation classification focused on distinguishing citations based on their functions or the author's reason for citing an article. However, newer classification methods characterizing citations based on their importance and influence were not introduced before 2015. Existing research in citation importance classification uses feature-based binary classification approaches. Two of the most prominent research works in this area were conducted by Zhu et al. (2015) and Valenzuela et al. (2015). Although the former identified 40 different features for detecting a subgroup of references from the bibliography that are influential to the citing document, the latter used 12 slightly overlapping features for characterizing both direct as well as indirect citations as *incidental* or *important*. Pride and Knoth (2017a, b) analyzed the features from the works mentioned above to identify the most prominent predictors for citation influence classification. By measuring the correlation between the earlier features and the truth label, they find abstract similarity to be the most predictive feature.

Table 4 illustrates some of the prominent literature in the area of citation importance classification. All the literature reviewed in this paper for citation importance identification use binary classification schemes; *Incidental*/*Nonimportant* and *Important*/*Influential*. The scheme developed by Valenzuela et al. (2015) considers citations belonging to the categories *Using* and *Extending the work* as *Important*, whereas the *Background* and *Comparison* related citations are treated as *Incidental*. The most widely used data set for this task is from Valenzuela et al. (2015), using the Association for Computational Linguistics (ACL) Anthology, containing 465 citation pairs. Qayyum and Afzal (2019) used two sets of data, one from Valenzuela et al. (2015), annotated by the domain experts, and a second corpus, which was annotated by the authors themselves. The distribution of class instances shows that less than 15% of citation contexts belong to the *Influential* or *Important* class for all studies. All the studies mentioned in this study used simple machine learning-based models such as Support Vector Machine (SVM), Logistic Regression (LR), k-Nearest Neighbors (kNN), etc., and the best performed classifier in most cases is Random Forest (RF). The most prominent predictor in all the cases is the number of times a paper is cited within the citing paper (Nazir, Asif et al., 2020b; Valenzuela et al., 2015; Wang et al., 2020b; Zhu et al., 2015).

4.3. Citation Function

Citations act as a link between the citing and the cited document, performing one of several functions. For instance, some citations indicate research that is foundational to the citing work, whereas others could be used for comparing, contradicting, or providing background information for the proposed work. Classification of citations according to their purpose serves several applications, with citation analysis for research evaluation being one of the key application areas (Dong & Schäfer, 2011; Jochim & Schütze, 2012). "Citation function reflects the specific purpose a citation plays with respect to the current paper's contributions" (Jurgens et al., 2018). The technique for identifying the citation function, however, requires the development

Table 4. Annotation schemes and data sets used for Citation Importance classification

Paper	Categories	Data Size	Important Findings
Zhu et al. (2015)	**Influential—10.3%**	100 papers	• Using authors themselves as annotators for identifying key references.
	Noninfluential—89.7%	3,143 citing paper–reference pairs	• Key predictors are reference count and similarity between cited title and core sections of citing paper.
Valenzuela et al. (2015)	**Incidental—85.4%** (1) Related work (2) Comparison	465 instances represented as (cited, citing paper) tuple	• Out of the total annotations, only 69 instances were present in the important category. • Identification of direct and indirect citations critical in citation importance classification.
	Important—14.6% (1) Using the work (2) Extending the work		
Qayyum and Afzal (2019)	**Important**	(1) Data set same as Valenzuela et al. (2015)	• The use of metadata alone produces good results, compared to methods employing content-based features.
	Nonimportant	(2) 488 paper-citation pairs from Computer Science	
Wang, Zhang et al. (2020b)	**Important**	(1) Data set same as Valenzuela et al. (2015)	• Citation intents such as Background and Methods were more effective in identifying important citations.
	Nonimportant	(2) 458 citation pairs on ACL Anthology	

of a classification schema, constituting the various functions under which citations in a research paper fall (Radoulov, 2008).

The earlier taxonomies largely inspired the recent developments in the citation classification. As an example, citation function classification strategy by Spiegel-Rösing (1977) was adapted later by several studies (Abu-Jbara et al., 2013; Jha et al., 2017; Teufel et al., 2006a, b). To find the relational information between the cited and the citing text, Teufel et al. (2006a) developed a taxonomy of 12 categories, inspired by Spiegel's scheme, where the four top-level classes captured the explicitly mentioned weakness, comparison or contrast, agreement/usage/compatibility with the cited research and finally a neutral category. Abu-Jbara et al. (2013) and Jha et al. (2017) experimented with more compressed categories containing six classes, namely, *Criticizing*, *Comparison*, *Use*, *Substantiating*, *Basis*, and *Neutral*. The earlier schema by Moravcsik and Murugesan (1975) was later studied using automated approaches by Dong and Schäfer (2011), Jochim and Schütze (2012), and Meng, Lu et al. (2017), where Dong and Schäfer and Meng et al. focused only on the *Organic vs. Perfunctory* dimension of the taxonomy. Jochim and Schütze (2012) noted that the "... most difficult facet for automatic classification ..." was *Confirmative vs. Negational* and the easiest was *Conceptual vs. Operational*. Bertin and Atanassova (2012) introduced a hierarchical classification scheme with a higher level containing five generic rhetorical categories and 11 specific classes at the lower level. The use of ontologies for describing the nature of citation is explored by Shotton (2010). The CiTO (Citation Typing Ontology)[8] captures the relationship between the citing and the cited articles and visualizes this information using Semantic Web technologies (RDF, OWL, etc.). A recent taxonomy introduced by scite_[9] classifies citation types into the classes: *Supporting*, *Disrupting*, and *Mentioning*, based on the level of evidence provided by citations.

4.4. Citation Polarity

Several studies concerning the development of citation classification taxonomies examine the polarity of the citation context as well for characterizing the cited articles. Abu-Jbara et al. (2013), Jha et al. (2017), Lauscher et al. (2017), Li, He et al. (2013), and Teufel et al. (2006a) included the categories Positive, Negative, and Neutral classes for capturing the sentiment associated with the citations. Li et al. (2013) proposed a two-level citation function schema, where the abstract top-level featured the sentiment classes and a lower set of categories capturing the fine-grained citation functions. The schema includes categories for representing the relation between two cited works and research breakthroughs in a field. Jha et al. (2017) differentiate citation function and polarity, where the former conveys the citer's motivation and the latter specifies the author's attitude towards the cited work. Teufel et al. (2006a, b) wrapped up the entire 12 categories as: Positive – *PMot*, *PUse*, *PBas*, *PModi*, *PSim*, *PSup*, Negative—*Weak*, *CoCo-*, and Neutral—*CoCoGM*, *CoCoR0*, *CoCoXY*, *Neut*, with the aim of performing sentiment analysis over the citations.

5. DATA SETS

In this section we discuss the common data sets for citation classification, the data source from which these corpora are derived, and finally the annotation procedures used by the authors for creating the data sets.

[8] https://purl.org/spar/cito
[9] https://scite.ai/

5.1. Data Sources

Tables 4 and 5 show the information related to the data set sources for citation importance and function classification respectively. Papers in Computer Science, specifically Computational Linguistics, have been a popular data source choice for citation classification tasks. This is largely attributed to the release of two prominent data sets for bibliographic research from ACL Anthology[10]: the ACL Anthology Reference Corpus (ACL ARC) (Bird, Dale et al., 2008) and the ACL Anthology Network (AAN) corpus (Radev, Muthukrishnan et al., 2013). The former consists of 10,921 articles, with full text and metadata extracted from the PDF files, and the latter is a networked citation database containing more than 19,000 NLP papers, with information about the paper citation, author citation, and author collaboration networks, besides the full text and metadata.

Another subject area of interest in the citation analysis research is the Biomedical domain. PubMed[11] and PubMed Central (PMC)[12], archives maintained by the U.S. National Institutes of Health (NIH) offers free access to the citation database, abstracts, and the full text corresponding to the biomedical and life sciences journal articles. Microsoft Academic Graph (MAG) (Sinha, Shen et al., 2015) is a heterogeneous graph that contain records of scholarly publications, citation relationships, bibliographic metadata, and the field of study. As opposed to Web of Science and Scopus, MAG also extracts citation context information, which is "... individual paragraphs immediately preceding each citation ..." (Wang, Shen et al., 2020a). However, by the end of 2021 Microsoft research will discontinue all MAG-related services. A new Semantic Scholar Open Research Corpus (S2ORC) (Lo, Wang et al., 2020), which is a large English language scientific data set, contains full text, metadata and citation links for 8.1 million open access publications. This data set is derived from sources such as PubMed and arXiv.

5.2. Annotated Data Sets

Table 5 shows the existing data sets for citation function classification. In an attempt to classify citations based on their rhetorical functions, Teufel et al. (2006a, b) developed a new data set[13] using 116 conference articles and 2,829 citation instances from Computational Linguistics tagged with citation functions. Another most widely used data set, developed by Abu-Jbara et al. (2013) contain annotations for citation purpose, polarity as well as information regarding the relatedness of sentence to the target citation. This AAN based data set was further studied extensively by Jha et al. (2017) and Lauscher et al. (2017)[14]. Jurgens et al. (2018) created a corpus with annotations for six citation functions using 585 papers from the ACL-ARC corpus[15]. The same data set was also used by authors for experiments related to analyzing the narrative structure of papers, venue evolution, and modeling the evolution of the NLP field.

To address the limitations caused by the nonavailability of larger annotated data sets, Cohan et al. (2019)[16] and Pride and Knoth (2020) introduced two new corpuses, SciCite and ACT, respectively. The former contains annotations for 11,020 instances of papers from Computer Science and Medicine and the later is a multidisciplinary data set with 11,233 instances obtained using full-text research papers from CORE. As with citation importance

[10] https://www.aclweb.org/anthology/
[11] https://pubmed.ncbi.nlm.nih.gov/
[12] https://www.ncbi.nlm.nih.gov/pmc/about/intro/
[13] https://www.cl.cam.ac.uk/~sht25/CFC.html
[14] https://clair.si.umich.edu/corpora/citation_sentiment_umich.tar.gz
[15] https://jurgens.people.si.umich.edu/citation-function/
[16] https://github.com/allenai/scicite

Table 5. Citation purpose and polarity classification schemes

Paper	Classification scheme	Data set	Important findings
Garzone and Mercer (2000)	(1) **Negational**—7 classes (2) **Affirmational**—5 classes (3) **Assumptive**—4 classes (4) **Tentative**—1 class (5) **Methodological**—5 classes (6) **Interpretational/Developmental**—3 classes (7) **Future Research**—1 class (8) **Use of Conceptual Material**—2 classes (9) **Contrastive**—2 classes (10) **Reader Alert**—4 classes	14 journal articles from Physics (8) and Biochemistry (6)	• Poor performance of classifier on unseen Physics articles (less well-structured), compared to Biochemistry articles (more well-structured)
Nanba et al. (2000)	(1) **Type B**—Basis (2) **Type C**—Comparison or Contrast (3) **Type O**—Other	395 papers in Computational Linguistics (e-print archive)	• Performance of the classifier solely depends on the cue phrases, absence of which causes wrong prediction
Pham and Hoffmann (2003)	(1) **Basis** (2) **Support** (3) **Limitation** (4) **Comparison**	482 citation contexts and 150 unseen citation contexts	• Incremental knowledge acquisition using the tool KAFTAN for citation classification

Table 5. (*continued*)

Paper	Classification scheme	Data set	Important findings
Teufel et al. (2006a, b)	(1) Weakness of cited approach—**Weak**—3.1% (2) Contrast/Comparison in Goals/Methods (neutral)—**CoCoGM**—3.9% (3) Contrast/Comparison in Results (neutral)—**CoCoR0**—0.8% (4) Unfavorable Contrast/ Comparison—**CoCo**—1.0% (5) Contrast between two cited methods—**CoCoXY**—2.9% (6) Author uses cited work as starting point—**PBas**—1.5% (7) Author uses tools/algorithms/ data—**PUse**—15.8% (8) Author adapts or modifies tools/algorithms/data—**PModi**—1.6% (9) Citation is positive about approach or problem addressed—**PMot**—2.2% (10) Author's work and cited work are similar—**PSim**—3.8% (11) Author's work and cited work are compatible/ provide support for each other—**PSup**—1.1% (12) Neutral description/not enough textual evidence/unlisted citation function—**Neut**—62.7%	116 articles and 2,829 citation instances from articles in Computational Linguistics (e-print archive)	• 60% of instances belong to neutral class • Low frequency of negative citations

Le et al. (2006)	(1) Paper is based on the cited work (2) Paper is a part of the cited work (3) Cited work supports this work (4) Paper points out problems or gaps in the cited work (5) Cited work is compared with the current work (6) Other citations	811 citing areas in 9000 papers from ACM Digital Library and Science Direct	• Use of finite-state machines for citation type recognition does not require domain experts or knowledge about cue phrases
Agarwal et al. (2010)	(1) **Background/Perfunctory** (2) **Contemporary**, (3) **Contrast/Conflict** (4) **Evaluation**, (5) **Explanation** (6) **Method**, (7) **Modality** (8) **Similarity/Consistency**	1,710 sentences from 43 open-access full text biomedical articles	• Model performed less on classes, Evaluation, Explanation & Similarity/Consistency • Infrequent keywords not recognized by model
Shotton (2010)	**Factual:** (1) cites, (2) citesAsAuthority, (3) isCitedBy, (4) citesAsMetadataDocument, (5) citesAsSourceDocument, (6) citesForInformation, (7) obtainsBackgroundFrom, (8) sharesAuthorsWith, (9) usesDataFrom, (10) usesMethodIn **Rhetorical—Positive:** (1) confirms, (2) credits, (3) updates, (4) extends, (5) obtainsSupportFrom, (6) supports **Rhetorical—Negative:** (1) corrects, (2) critiques, (3) disagreesWith, (4) qualifies, (5) refutes	Ontology developed for Biomedical articles	• OWL-based tool, CiTO for characterizing the nature of citations

Table 5. (*continued*)

Paper	Classification scheme	Data set	Important findings
	Rhetorical—Neutral: (1) discusses, (2) reviews		
Dong and Schäfer (2011)	(1) **Background**—65.04% (2) **Fundamental idea**—23.80% (3) **Technical basis**—7.18% (4) **Comparison**—3.95%	1768 instances & 122 papers from ACL Anthology (2007 and 2008)	• Use of Ensemble-style self-training reduces the manual annotation work
Jochim and Schütze (2012)	(1) **Conceptual**—89.2% vs. **Operational**—10.8% (2) **Organic**—10.1% vs. **Perfunctory**—89.9% (3) **Evolutionary**—89.8% vs. **Juxtapositional**—10.2% (4) **Confirmative**—91.4% vs. **Negational**—8.6%	84 papers and 2008 citation from papers in 2004 ACL Proceedings (ARC)	• Annotation of four facets using Moravscik's scheme instead of a single label
Abu-Jbara et al. (2013)	**Purpose:** (1) **Criticizing**—14.7% (2) **Comparison**—8.5% (3) **Use**—17.7% (4) **Substantiating**—7% (5) **Basis**—5% (6) **Neutral**—47% **Polarity:** (1) **Positive**—30% (2) **Negative**—12% (3) **Neutral**—58%	3,271 instances from 30 papers in ACL Anthology Network (AAN)	• 47% of citations belong to the class Neutral • Citation Purpose classification Macro-Fscore: 58.0%
Xu et al. (2013)	(1) **Functional**—48.4% (2) **Perfunctory**—50% (3) **Fallback**—1.6%	ACL Anthology Network corpus (AAN)	• Self-citations are skewed to the class Functional • Authors citing more has more functional citations

Li et al. (2013)	(1) **Based on**—2.8% (2) **Corroboration**—3.6% (3) **Discover**—12.3% (4) **Positive**—0.1% (5) **Practical**—1% (6) **Significant**—0.6% (7) **Standard**—0.2% (8) **Supply**—1.2% (9) **Contrast**—0.6% (10) **Cocitation**—33.3% (11) **Neutral**, (12) **Negative**—(Omitted both these categories)	91 Biomedical articles and 6,355 citation instances from Biomedical articles (PubMed)	• Coarse-grained sentiment classification performs only slightly better than fine-grained citation function classification
Hernández-Álvarez et al. (2016)	**Purpose:** (1) **Use**—(a) Based on, Supply—16.1% (b) Useful—33.7% (2) **Background**—(c) Acknowledge/ Corroboration/Debate—37.4% (3) **Comparison**—(d) Contrast—5.3% (4) **Critique**—(e)Weakness—6% (f) Hedges—1.8% **Polarity:** (1) **Positive**—28.7% (2) **Negative**—9.7%, (3) **Neutral**—64.7%	2,092 citations in 85 papers from ACL Anthology Network (AAN)	• Classes Acknowledge and Useful dominate the data distribution for purpose classification • Neutral class has more than 50% of instances

Table 5. (*continued*)

Paper	Classification scheme	Data set	Important findings
Munkhdalai, Lalor, and Yu (2016)	**Function:** (1) **Background**—30.5%, 20.5% (2) **Method**—23.9%, 18.2% (3) **Results/findings**—45.3%, 38.3% (4) **Don't know**—0.1%, 0.06% **Polarity:** (1) **Negational**—4.8%, 2.6% (2) **Confirmative**—75%, 59.8% (3) **Neutral**—19.8%, 19% (4) **Don't know**—0.2%,0.1%	Data 1—3,422 (Function), 3,624 (Polarity) citations Data 2—4,426(Function), 4,423(Polarity) citations from 2,500 randomly selected PubMed Central articles	• Majority of citations annotated as results and findings • Bias of citations towards positive statements
Fisas et al. (2016)	(1) **Criticism**—23%: (a) Weakness, (b) Strength, (c) Evaluation, (d) Other (2) **Comparison**—9%: (a) Similarity, (b) Difference (3) **Use**—11%: (a) Method, (b) Data, (c) Tool, (d) Other (4) **Substantiation**—1% (5) **Basis**—5%: (a) Previous own Work, (b) Others work, (c) Future Work (6) **Neutral**—53%: (a) Description, (b) Ref. for more information, (c) Common Practices, (d) Other	10,780 sentences from 40 papers in Computer Graphics	• A multilayered corpus with sentences annotated for (1) Citation purpose, (2) features to detect scientific discourse and (3) Relevance for summary
Jha et al. (2017)	Same as Abu-Jbara et al. (2013)	3500 citations in 30 papers from ACL Anthology Network (AAN)	• Developed data sets for reference scope detection and citation context detection • Comprehensive study aimed at applications of citation classification

Lauscher et al. (2017)	Same as Abu-Jbara et al. (2013)	Data sets from Abu-Jbara et al. (2013) and Jha et al. (2017)	• Heavy skewness of data set towards less informative classes for both schemes • Use of domain-specific embeddings does not enhance results
Jurgens et al. (2018)	(1) **Background**—51.8% (2) **Uses**—18.5% (3) **Compares or Contrasts**—17.5% (4) **Motivation**—4.9% (5) **Continuation**—3.7% (6) **Future**—3.6%	1,969 instances from ACL-Anthology Reference Corpus (ACL-ARC)	• Majority of instances belong to class Background • Error analysis shows the importance of citation context identification for result improvement
Su, Prasad et al. (2019)	(1) **Weakness**—2.2% (2) **Compare and Contrast**—6.6% (3) **Positive**—20.6% (4) **Neutral**—70.6%	ACL-ARC Computational Linguistics	• Highly skewed data set with majority of instances belonging to Neutral class • Use of Multitask learning for citation function and provenance detection
Cohan et al. (2019)	(1) **Background**—58% (2) **Method**—29% (3) **Result Comparison**—13%	6,627 papers and 11,020 instances from Semantic Scholar (Computer Science & Medicine)	• Introduction of new data set known as SciCite • The best state-of-the-art macro-fscore obtained using BiLSTM attention with ELMO vector & structural scaffolds
Pride, Knoth, and Harag (2019)	(1) **Background**—54.61% (2) **Uses**—15.51% (3) **Compares/Contrasts**—12.05% (4) **Motivation**—9.92% (5) **Extension**—6.22%, (6) **Future**—1.7%	Multidisciplinary data set of 11,233 instances from CORE	• Largest multidisciplinary author annotated data set

classification, the commonly used data set, released by Valenzuela et al. (2015), with citations in the form of 465 tuples (cited paper, citing paper) and annotations for both citation importance and type, is shown in Table 4.

5.3. Annotation Guidelines

Annotation guidelines describe the criteria required by the citations to qualify for each category. Teufel et al. (2006a) used annotation guidelines that stated the requirement for annotating only single "... *explicitly signalled citation functions* ...." The developers of the SciCite data set used 50 test questions annnotated by domain experts in an effort to disqualify annotators whose annotation accuracy was lesser than 50% (Cohan et al., 2019). The authors also used a fourth class, *Others,* besides the original three classes, to improve the annotation quality. Abu-Jbara et al. (2013) sought for three different tags from the annotators: Sentences relevant to citation, Citation Purpose, and Citation Polarity. The number of annotators ranges from two to multiple people. Annotators in most cases are domain experts or graduate students with a background in the subject (Bakhti et al., 2018; Fisas et al., 2016; Hernández-Álvarez et al., 2017; Jha et al., 2017). The work of Pride and Knoth (2020), however, differs from other annotation works by employing authors themselves as annotators based on the assumption that they are most qualified to decide what they meant by each citation they used in their manuscript.

To make the annotation process easier, specialized tools are used in certain cases. For example, Jurgens et al. (2018) employed the Brat rapid annotation tool[17] and two NLP experts for doubly annotating citations. Fisas et al. (2016), Jochim and Schütze (2012), Pride et al. (2019), Radoulov (2008), and Teufel et al. (2006a) developed web-based annotation tools for simplifying the task. To compute the agreement between the annotators, measures such as the Kappa coefficient (Abu-Jbara et al., 2013; Agarwal et al., 2010; Dong & Schäfer, 2011; Teufel et al., 2006a), Cohen's Kappa coefficient, the Krippendorff coefficient (Hernández-Álvarez et al., 2017) and other confidence scores (Cohan et al., 2019) are utilized. Citation annotations by independent annotators is a difficult task because often authors do not always state their intentions for citing explicitly (Gilbert, 1977; Teufel et al., 2006a; Zhu et al., 2015). Alternatively, the developers of the citation schema (Agarwal et al., 2010; Teufel et al., 2006a) or the cited authors themselves annotated the citations (Nazir et al., 2020b; Pride et al., 2019; Zhu et al., 2015). Recently, crowdsourcing platforms have also been utilized for tagging citation labels (Cohan et al., 2019; Munkhdalai et al., 2016; Pride et al., 2019; Su et al., 2019).

6. PREPROCESSING

Text preprocessing is typically applied prior to undertaking citation function and importance classification. The process typically involves extracting text from documents (most commonly PDFs), parsing the contents for extracting metadata, references, citation context, etc. and finally preparing the text for feature extraction. The general prototypical architecture for citation classification is illustrated in Figure 4. In this section, we provide an overview of scientific document parsing, the tools used, and the methods for citation context detection.

6.1. Document Parsing

The initial step in citation classification involves parsing of the PDF files for reference extraction and citation context detection. First, the bibliographic section of the PDF file is identified, followed by the extraction of reference strings. Reference parsing open source systems based on Conditional Random Field (CRF) such as ParsCit (Councill, Giles, & Kan, 2008), GROBID

[17] https://brat.nlplab.org/

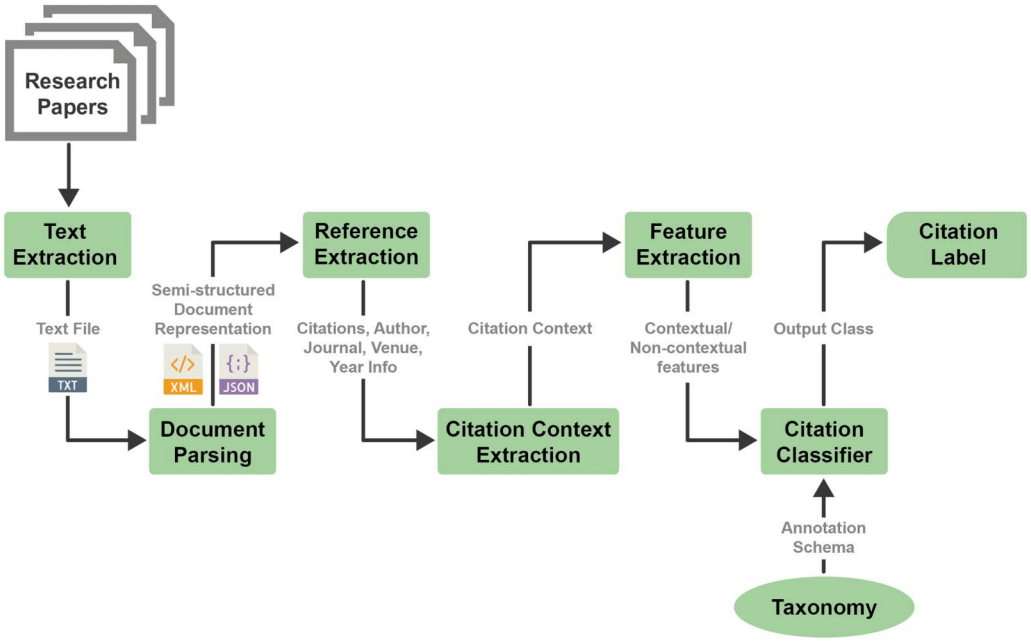

Figure 4. Prototypical diagram for citation classification.

(Lopez, 2009), CERMINE (Tkaczyk, Szostek et al., 2015) and Science Parse[18] aim at converting the plain text or PDFs to a more semistructured format such as XML/JSON for extracting not only the metadata but also other information corresponding to the abstract, sections, etc. from the scholarly articles. ParsCit processes the reference string and extracts the citation context and the following 13 fields from the bibliography:

(1) Author
(2) Book title
(3) Date
(4) Editor
(5) Institution
(6) Journal
(7) Location
(8) Note
(9) Pages
(10) Publisher
(11) Tech
(12) Title
(13) Volume

Unlike ParsCit, which accepts the input data only in the UTF-encoded text format, GROBID, CERMINE, and Science Parse are capable of directly processing the PDF files. Other tools for extracting the in-text citations are PDFX (Constantin, Pettifer, & Voronkov, 2013), Crossref pdfextract[19], and Neural ParsCit (Prasad, Kaur, & Kan, 2018), where the former two are rule-based and the later employs Long Short Term Memory (LSTM) neural networks.

[18] https://github.com/allenai/science-parse
[19] https://github.com/Crossref/pdfextract

6.2. Citation Context Detection

Authors may use citations to substantiate or refute their claims. The citation context, which contains the pointer to the referenced article reflects the information about the cited paper (Su et al., 2019). Abu-Jbara et al. (2013) and Jha et al. (2017) defined explicit citing sentences as the "... sentences in which actual citations appear ...." Research papers at times include sentences devoid of any citation that is related to the cited article. Such extended context, constituting sentences with indirect and implicit references to the cited paper surrounding the citing sentence, are also studied for improved citation classifier performance (Athar & Teufel, 2012b; Hernández-Álvarez & Gomez, 2016). Rotondi, Di Iorio, and Limpens (2018) argue the need for considering the subject domain and the specificity of the language before choosing the citation context width. Detecting the citation context is an importance step as this is considered a prerequisite for citation classification (Lauscher et al., 2017; Rotondi et al., 2018).

Finding the optimal window size for citation context is critical, as this area determines the amount of information processed for successful identification of the citation class. Often this could be challenging as there are considerable variations in the amount of text surrounding the citations that talk about the cited paper. Rotondi et al. (2018) mention the following possibilities for citation context window size: Fixed number of characters—use of 200 characters by ParsCit[20], (Jurgens et al., 2018); Citing sentence—(Bertin, Atanassova et al., 2016; Cohan et al., 2019; Garzone & Mercer, 2000; Hassan, Safder et al., 2018; Pride et al., 2019; Sula & Miller, 2014; Valenzuela et al., 2015); and Extended context—three or more sentences including the sentences immediately preceding and following the citing sentence (fixed context) (Abu-Jbara et al., 2013; Agarwal et al., 2010; Athar & Teufel, 2012a; Hernández-Álvarez et al., 2017; Munkhdalai et al., 2016; Nanba et al., 2000; Su et al., 2019; Teufel et al., 2006a) and using all mentions of citations in the article (adaptive context) (Athar & Teufel, 2012b).

The usability of extended context for performance improvement has always encountered the following two concerns among the researchers: the introduction of noise while incorporating additional context (Cohan et al., 2019) and the loss of information in the case of using just the citing sentence for citation classification (Athar & Teufel, 2012b). Abu-Jbara et al. (2013) use a sequence labeling technique for identifying the citation context. The authors found that a window size of four sentences often contained the related context, one sentence before the citing sentence, the citing sentence itself and two sentences after the citing sentence. Valenzuela et al. (2015) and Xu et al. (2013) claim to obtain the same level of performance as that of the classifier with extended context by using the citing sentence alone. However, earlier studies related to citation sentiment demonstrate that the polarity and author's attitude, in the form of hedging, are most likely to be found outside the citing sentence (Athar & Teufel, 2012b; Di Marco et al., 2006).

6.3. Mitigating Data Set Skewness

A major problem concerning the citation classifiers' performance issues is attributed to the highly skewed nature of the classes. Several data sets report a higher number of instances for the nonimportant citation types such as *Background* or *Neutral* and a relatively lower number of cases for more important categories such as *Extension* or *Future*. Dong and Schäfer (2011) reduced the original corpus with class distribution ratio from 16:6:1.8:1 to 5:2.5:2:1 for the classes *Background*, *Fundamental Idea*, *Technical Basis*, and *Comparison*, respectively, to obtain a more balanced data set. The use of category-specific annotations for increasing the

[20] https://parscit.comp.nus.edu.sg/

number of instances in the rare classes is also employed to mitigate the class-imbalance problem (Jurgens et al., 2018; Li et al., 2013; Zafar, Ahmed, & Islam, 2019). Jurgens et al. (2018), Nazir et al. (2020b), and Qayyum and Afzal (2019) applied SMOTE to create synthetic instances to tackle the skewness in the data set. Zhu et al. (2015) down-sampled the *non-influential* instances during cross-validation to make it the same as that of the *influential* citations. Another approach is the removal of categories that do not convey any information. Abu-Jbara et al. (2013) eliminated the class *Neutral*, which contains more than 50% of the total number of instances and performs a binary classification for polarity detection to obtain more intuitive results. Analyzing the SciCite data set, Pride and Knoth (2020) found that authors used an oversampling technique on the underrepresented *Methods* class in the data set.

7. FEATURES FOR CITATION CLASSIFICATION

Automatic citation classification based on machine learning methods makes use of features that help capture the relationship between the citing and the cited papers. The features are manually determined and the text-based citation context is analyzed for extracting informative signals. Tables 6 and 7 illustrate the features used by some of the literature related to citation function and importance classification. The classification of citations in the existing literature takes into account the following different feature dimensions.

7.1. Contextual Features

The contextual features are categorized at a higher level as Syntactic and Semantic, according to how and why the citations are described in the text. The latter is further classified as Textual-based, Similarity-based, and Polarity-based.

7.1.1. Syntactic features

The use of dependency relations was found to be an effective signal for capturing the syntactic information from the citation context (Dong & Schäfer, 2011; Jochim & Schütze, 2012; Li et al., 2013; Meng et al., 2017). Bertin and Atanassova (2014) and Bertin et al. (2016) emphasize the importance of verbs in understanding the nature of the relation between the citing and the cited articles. Dong and Schäfer (2011) reported the best results for an ensemble classifier using the syntactic POS tag features specific to each class. The application of syntactic features alone resulted in performance improvement compared to the baseline model for Jochim and Schütze (2012) and Li et al. (2013). Teufel et al. (2006b) used verb tense and voice for identifying citation contexts corresponding to previous work, future work, and work performed in the citing paper. Jha et al. (2017) showed that the features having direct dependency relation to the cited paper, for instance, closest verb, adjective, adverb, and subjective cue, are the most promising signals.

7.1.2. Semantic features

The application of metadiscourse or cue words/phrases for automatic citation classification has been extensively studied in the past (Dong & Schäfer, 2011; Jurgens et al., 2018; Mercer & Di Marco, 2003; Teufel et al., 2006b; Xu et al., 2013). Mercer and Di Marco (2003) acknowledge the relevance of cue words as a "... conjunction or connective that assists in building the coherence and cohesion of a text ...." The authors studied the occurrence of cue phrases in the full-text IMRaD (Introduction, Method, Result and Discussion) sections and citing sentence as well as in the citation context and came to the conclusion about the significant presence of discourse cues in citation context, which makes these critical determiners for categorizing citations based on their roles. The presence of hedging cue words

Table 6. Features used for citation function classification

	Features used						
	Contextual				Noncontextual		
		Semantic					
Papers	Syntactic	Textual-Based	Similarity-Based	Polarity-Based	Positional-Based	Frequency-Based	Other
Teufel et al. (2006b)	• Verb Tense • Voice • Modality	• Cue phrases			• Location within (1) Article, (2) Paragraph, (3) Section		• Self-citation
Dong and Schäfer (2011)	• POS Tags	• Cue Words specific to classes			• Location within section	• Popularity • Density • Avg Density	
Athar (2011)	• POS Tags • Dependency Relations	• n-grams (n = 1–3) • Subjectivity cues • Negation		• Scientific polarity lexicon		• Number of (1) Adjectives, (2) Adverbs, (3) Pronouns, (4) Modals, (5) Cardinals, (6) Negation phrases, (7) Valance shifters	• Name of the primary author • Sentence splitting

Jochim and Schütze (2012)	• Dependency Relations • POS Tag patterns • Citation is a constituent • Author linked to comparative • Citation linked to comparative • Citation is in contrastive clause • Author linked to positive sentiment • Same as Teufel et al. (2006b) • Sentence has modal verb • Dependency root node • Main verb • First person POS • Third person POS • Comparative/ superlative POS • Has "but" • Has "cf."	• Cue Words • n-grams (n = 1–3)	• Scientific polarity lexicon • General polarity lexicon • General positive lexicon • General negative lexicon	• Section • Location within (1) Paper (2) Paragraph (3) Section (4) Sentence	• Popularity • Density • Avg Density	• Self-citation • Has resource • Has tool	

Table 6. (*continued*)

	Features used						
	Contextual				Noncontextual		
		Semantic					
Papers	Syntactic	Textual-Based	Similarity-Based	Polarity-Based	Positional-Based	Frequency-Based	Other
Xu et al. (2013)	• Whether citations used in parenthesis	• Cue patterns • n-grams (n = 1–3)			• Location within paper	• Number of citation anchors within sentence.	• Author relationships • Paper relationships • Centrality measures • Self-citations
Li et al. (2013)	• 3rd person pronoun • POS Tags • Dependency Relations	• n-grams • Cue word/phrases					• Presence of formula, graph and table in citation context
Abu-Jbara et al. (2013), Jha et al. (2017)	• Closest Verb/ Adjective/ Adverb • Contains 1st/3rd person pronoun • Dependency Relations	• Negation, Speculation, Closest Subjectivity Cue • Contrary Expressions			• Section	• Reference Count	• Is Separate • Self Citation
Bakhti et al. (2018)		• n-grams (n = 2–3) • Cue phrases					

Jurgens et al. (2018)	• Verb Tense • Lengths of sentence and clause • Bootstrapped and Custom function patterns • Used with Parenthesis • Citation prototypicality • Whether used in nominative/ parenthetical form • Whether preceded by (1) Pascal-cased word, (2) All-capital case word	• Extended Cue phrases (Teufel et al., 2006b) • Citation context topics	• Topical similarity with cited paper	• Location within (1) Paper (2) Section (3) Subsection (4) Sentence (5) Clause • Canonicalized section title	• Direct Citations • Direct & Indirect citations/section type • Indirect Citations • Fraction of bibliography used by reference • Citation in (1) Subsection, (2) Sentence, (3) Clause • Common Citations count	• Self-citation • Year difference in publication dates • Citing paper's venue • Reference's venue • Reference's citation count & PageRank • Reference's Hub & Authority scores & Network Centrality	

Table 7. Features used for citation importance classification

	Features used						
	Contextual				Noncontextual		
		Semantic					
Papers	Syntactic	Textual-Based	Similarity-Based	Polarity-Based	Positional-Based	Frequency-Based	Other
Zhu et al. (2015)	• Explicit reference of cited author • Whether citations (1) Appear alone (2) Appear first in the list	• Cue words for determining Cited article's (1) Relevance (2) Recentness (3) Extremeness (4) Degree of Comparison * # of (1) Strong & (2) Active words • Word-net features • General Inquirer features	• Similarity between Cited Title and (1) Title, (2) Abstract, (3) Introduction, (4) Conclusion, & (5) Core sections • Similarity between citation context and (1) Title, (2) Abstract, (3) Introduction, (4) Conclusion	• # of positive words in citation context • Emotion Lexicon for detecting (1) Sentiment and (2) Emotive words	• Whether citations appear at the (1) Beginning or (2) End of the sentence • Position of citing sentence based on (1) Mean, (2) Standard variance, (3) First, (4) Last	• Citation counts in (1) Entire Paper, (2) Introduction, (3) Related Work, (4) Core Sections * # of sections where reference appears • # Global citations	• Self-citations • Publication year
Valenzuela et al. (2015)		• Citation considered helpful based on cue phrases	• Similarity between abstracts		• Citation appears in table or caption	• # Direct citations • # Direct citations per section • # Indirect citations • # Indirect citations per section • 1/# of references • # of paper citations/all citations • # of total citing papers after transitive closure	• Author overlap • PageRank • Field of cited paper

Hassan et al. (2017, 2018)		• Cue words for (1) Related Work (2) Comparative citations, (3) Using & (4) Extending current work	• Similarity between citing text and cited abstract	• Citations in sections (1) Introduction (2) Literature Review (3) Method (4) Experiment (5) Discussion (6) Conclusion	* # citation count for reference • # of citations from citing to cited paper	• Author Overlap
Qayyum and Afzal (2019)		• Cue words	• n-gram similarity and dissimilarity between titles (n = 1–3) • Ratio of keywords similarity to dissimilarity between pairs • Abstract similarity			• Author Overlap • Bibliographically coupled references
Nazir et al. (2020a)					• Citation frequency	
Nazir et al. (2020b)	• Section-wise weights for in-text citations		• Similarity score		• Citation frequency	

Table 7. (*continued*)

	Features used						
	Contextual				Noncontextual		
		Semantic					
Papers	**Syntactic**	**Textual-Based**	**Similarity-Based**	**Polarity-Based**	**Positional-Based**	**Frequency-Based**	**Other**
Wang et al. (2020b)			• Textual Similarity			• # of citations * # citations per year # citations in (1) Introduction, (2) Literature Review, (3) Method, (4) Conclusion, (5) Experiment, (6) Discussion • Mentioned frequency • # (1) Method, (2) Background, (3) Result extension citations	• Time Distance • Author Overlap • Total citation length • Average citation length • Maximum citation length

or phrases such as "Although," "would," "might," "is consistent with," and so forth, which captures the lack of certainty in citation contexts was noted by Di Marco et al. (2006). Jurgens et al. (2018) noted the presence of citation context topics and word vectors in the top 100 highest weighted features providing accurate information.

Other commonly used semantic features include similarity-based indicators. Hassan et al. (2017, 2018) and Pride and Knoth (2017a) operationalize these by measuring the semantic similarity between the cited abstract and the citing text using cosine similarity. They find this to be the best informative feature for citation importance classification. Similarly, for Zhu et al. (2015), the Pearson correlation coefficient between the features and the gold label indicates the effectiveness of the similarity-based features computed between the title/context of the cited paper with the different aspects of the citing paper. Popular deep learning approaches for citation classification rely on word representations such as Global Vectors for Word Representation (GloVe), Embeddings from Language Models (ELMo), and Bidirectional Encoder Representations from Transformers (BERT) for capturing the semantics from citation contexts (Beltagy, Lo, & Cohan, 2019; Cohan et al., 2019; Perier-Camby, Bertin et al., 2019).

Citation classification schemes with categories distinguishing the author's sentiment towards the cited article also use contextual features based on polarity. Abu-Jbara et al. (2013) and Jha et al. (2017) noted the importance of the cue phrases pertaining to subjectivity in classifying the citation polarity. The use of a lexicon based on scientifically polar words was explored by Athar (2011) and Jochim and Schütze (2012). Jochim and Schütze (2012) also used general-purpose polarity and positive and negative lexicons in their experiments, finding improvement in the performance of the classifier in identifying the facets, *Confirmative* vs. *Negational* as well as the *Evolutionary* vs. *Juxtapositional*.

7.2. Noncontextual Features

We categorize any extratextual features under this group as follows:

7.2.1. Positional-Based

The most common structural feature explored by the existing research relates to the location of the citations with respect to the document (Jochim & Schütze, 2012; Jurgens et al., 2018; Teufel et al., 2006b; Xu et al., 2013). The location of citations includes position with respect to the paper, paragraph, section, subsection, and sentence. Jurgens et al. (2018) added structural features corresponding to the relative citation position even in clauses. Bertin and Atanassova (2014) and Bertin et al. (2016) studied the in-document citation locations corresponding to the IMRaD structure of the document and came to the conclusion that the highly cited papers occur more frequently at the sections Introduction and Literature Review.

7.2.2. Frequency-Based

Abu-Jbara et al. (2013) and Jha et al. (2017) reported the number of citations in the context to be the most useful feature for identifying the citation purpose. Valenzuela et al. (2015) and Jurgens et al. (2018) added the number of direct and indirect citation counts in the features set. Both Dong and Schäfer (2011) and Jochim and Schütze (2012) take into account the different reference count aspects such as popularity (citations in the same sentence), density (citations in the same context) and average density (average density of neighboring sentences). The number of citations per section was found to be more correlated in deciding the academic influence by Zhu et al. (2015) and Wang et al. (2020b).

7.2.3. Other features

The most frequent miscellaneous feature used by the researchers is self-citation, which is an indication of whether any of the citing authors coauthored the cited paper (Abu-Jbara et al., 2013; Jha et al., 2017; Jochim & Schütze, 2012; Jurgens et al., 2018; Teufel et al., 2006b; Zhu et al., 2015). Xu et al. (2013) identified that self-citations are prominent in the class *functional*, which suggests that authors' new research is built on their previous work. Network-based features such as author relationships, paper relationships, citing paper/cited paper venue, and publication dates were also used for capturing the global information to classify citations (Hassan et al., 2017; Jurgens et al., 2018; Valenzuela et al., 2015; Xu et al., 2013).

8. AUTOMATIC CITATION CLASSIFICATION

Earlier citation classification methods mainly relied on the manual examination of citation context to identify citation types. To surpass the shortcomings of prior approaches, attempts were made to automate the process. The following sections discuss the existing automatic citation classification methods.

8.1. Rule-Based Methods

Garzone and Mercer (2000) introduced the first automated rule-based method, where the authors categorized citing sentences using 195 lexical matching and 14 parsing rules. A similar rule-based approach was later studied by Nanba et al. (2000) and Pham and Hoffmann (2003), where the former employed cue phrases for identifying the citing area and the latter devised a knowledge-acquisition system using Ripple Down Rules. These rule-based systems for classification suffer several downsides, including the requirement of a domain expert for developing the parsing rules and the identification of cue words specific to each citation type, which is a time-consuming process (Radoulov, 2008).

8.2. Traditional Machine-Learning-Based Methods

The first automatic machine learning-based citation classification approach was proposed by Teufel et al. (2006b). The authors obtained the best classification results using the IBk algorithm (a form of kNN). The authors also tested the classifier on the three polarity classes and attained a higher macro-f score of 0.71. Similar feature-based supervised learning techniques for citation classification were employed by several studies, which applied SVM (Bakhti et al., 2018; Hassan et al., 2017; Hernández-Álvarez et al., 2017; Jha et al., 2017; Meng et al., 2017; Xu et al., 2013; Zhu et al., 2015), RF (Jurgens et al., 2018; Pride & Knoth, 2017a; Valenzuela et al., 2015), Naive Bayes (NB) (Abu-Jbara et al., 2013; Agarwal et al., 2010; Dong & Schäfer, 2011; Sula & Miller, 2014), Maximum Entropy (MaxEnt) (Jochim & Schütze, 2012) and so forth for training the model.

Unlike the usual supervised learning approaches, Dong and Schäfer (2011) used a semisupervised ensemble learning model in an attempt to reduce the manual annotation of training data. The authors used a self-training algorithm to extend the training data set by using the predictions from the algorithm as labels for the unlabeled data set. Le et al. (2006) classified citation types using finite-state machines based on Hidden Markov Models (HMMs) and Maximum-Entropy Markov Models (MEMMs) to estimate the likelihood of each class. Radoulov (2008) also explored the possibility of applying semisupervised methods, where the authors first trained the model using NB on a small data set and later expanded the training set using an Expectation-Maximization (EM) algorithm.

A major shortcoming of the automatic citation classification based on machine learning methods is its requirement for manual determination of the features prior to training the model (Su et al., 2019). The success of such models relies on how well these features capture the syntactic as well as the semantic information from the citation context. Moreover, the citation classifiers are tested on smaller data sets due to the unavailability of larger corpora until 2019. Nevertheless, machine learning models are capable of producing acceptable results even with smaller training sets. Also, pattern-based features can still capture the properties of even the minority classes (Perier-Camby et al., 2019).

8.3. Deep-Learning-Based Methods

Recent years have witnessed the application of deep learning techniques for citation classification because of the progress in the field for solving NLP-related problems. Although sophisticated, the primary motivation for using neural architectures is their ability to identify features automatically, removing the pain of defining handcrafted features before classification. Perier-Camby et al. (2019) compared the performance of Bi-attentive Classification Network (BCN) and ELMo with the feature-based machine learning approach on the ACL-ARC data set. The authors emphasize the need for larger data sets for improved classification performance for deep learning methods. A combined model using Convolutional Neural Networks (CNN) and LSTM for capturing the n-grams and the long-term dependencies for multitask citation function and sentiment analysis was proposed by Yousif et al. (2019). A multitask learning approach using Cohan et al. (2019) identified the citation intent from the structural information, obtained using two auxiliary tasks: citation worthiness and section title, with the help of a bidirectional LSTM and attention mechanism, along with the ELMo vectors. A new transformer based model using BERT architecture, trained on 1.14 million scientific publications and called SciBERT, was developed by Beltagy et al. (2019). A larger SciBERT model, called S2ORC-SciBERT (Lo et al., 2020) is trained using a new corpus consisting of 8.1 million open access full-text scholarly publications.

9. EVALUATION METHODS

Table 8 shows the evaluation metric and the scores obtained on the most common data sets for citation classification. The frequently used evaluation method is macro averaged F-score because of the highly skewed nature of the data sets and the fact that macro averaging treats each category as a single entity, irrespective of the number of instances present in the class (Meng et al., 2017; Teufel et al., 2006b). The scores obtained for classification schemes with fine-grained categories often tend to be lower than the low-granularity schemes. Underrepresented categories of the fine-grained schemes reduce the overall macro F-score value (Perier-Camby et al., 2019). Similarly, the error analysis on the developed citation function classification model shows the increase in false positive rates for the dominating categories (Cohan et al., 2019). Because all evaluation scores mentioned in Table 8 are obtained under different settings of annotation schemes, classifiers, and data sets, a comparison of methods is nearly impossible.

10. SHARED TASKS

Recent years have witnessed the increasing popularity of shared tasks, usually organized as part of conferences or workshops. The intention here is to allow research improvements in the underresearched or underresourced areas of NLP, thus making possible the comparison of competing systems in such competitions (Nissim, Abzianidze et al., 2017). Although

Table 8. Evaluation scores obtained for existing citation classification data sets

Data set	# Instances	Classifier	Task	# classes	Metric	Score
Teufel et al. (2006b)	2,829	kNN (k = 3)	Purpose	12	Macro-F	0.57
					Kappa	0.57
				4	Macro-F	0.68
					Kappa	0.59
			Polarity	3	Macro-F	0.71
					Kappa	0.58
Dong and Schäfer (2011)	1,768	NB	Purpose	4	Macro-F	0.66
		SVM				0.79
Li et al. (2013)	6,355	MaxEnt	Purpose	11	F-Score	0.67
Abu-Jbara et al. (2013)	3,271	SVM	Purpose	6	Macro-F	0.58
					Accuracy	0.70
			Polarity	3	Macro-F	0.71
					Accuracy	0.81
		CNN + Multidisciplinary embedding	Purpose	6	F-Score	0.79
			Polarity	3	F-Score	0.82
Hernández-Álvarez et al. (2017)	2,120	SVM	Purpose	8	F-score	0.89
					ROC Area	0.95
			Polarity	3	F-score	0.93
					ROC Area	0.93
	348		Importance	3	F-score	0.94
Jurgens et al. (2018)	3,083	RF	Purpose	6	Macro-F	0.53
Cohan et al. (2019)	11,020	biLSTM Attention + ELMO & structural scaffolds	Purpose	3	Macro-F	0.84
		SciBERT	Purpose	3	Macro-F	0.85
Zhu et al. (2015)	3,143	NB	Importance	2	Macro-F	0.42
Valenzuela et al. (2015)	450	SVM	Importance	2	Precision	0.65
					Recall	0.90

research into the citation function has made considerable progress since the late 1970s, using a shared task as a benchmark for the future research in this direction has only recently been explored. Two shared tasks with regard to citation relevance and function classification were organized in 2020, the Microsoft Research—Citation Intent Recognition task and the 3C Citation Context Classification task.

10.1. Microsoft Research—Citation Intent Recognition

The shared task, Citation Intent Recognition, organized by Microsoft research as part of the WSDM Cup 2020[21] is an information retrieval task. The focus of this task is to separate the relevant citations from the superfluous ones. Given a paragraph or sentences containing citations, the participants were required to identify and retrieve the top three papers based on their relevance from a database. Using the description text as query, the participating teams should be able to retrieve the candidate papers from a pool of over 800,000 papers. The submitted systems were evaluated using Mean Average Precision @3 (MAP @3). The best information retrieval approach used BERT and LightGBM (Light Gradient Boosting Machine)[22] for the task (Chen, Liu et al., 2020). This shared task was hosted on the data science competition hosting platform, Biendata[23].

10.2. 3C Citation Context Classification Task

The 3C citation context classification task (Kunnath et al., 2020) organized by The Open University, UK as part of the workshop, WOSP 2020[24] and collocated with JCDL 2020[25], was the first shared task featuring the classification of citations based on its purpose and influence. This task utilized a portion (3,000 training instances) of the new multidisciplinary ACT data set (Pride et al., 2019), the largest data set annotated by authors themselves. The 3C shared task was organized as two subtasks: Subtask A—Citation Context Classification based on purpose[26], a multiclass classification problem based on the citation functions and Subtask B—Citation Context Classification based on influence[27], a binary task focusing on the citation importance classification. Both these subtasks were hosted as separate competitions using the Kaggle InClass competitions[28].

Subtask A involved the classification of citation into one of the following six classes based on the purpose: BACKGROUND, USES, COMPARES_CONTRASTS, MOTIVATION, EXTENSION, and FUTURE. The second classification subtask had the categories INCIDENTAL and INFLUENTIAL. Four teams participated in this shared task, of which three teams competed in both the tasks. All systems submitted were evaluated using a macro averaged F-score on a test set of 1,000 instances. Despite the recent advances in deep learning technologies, this shared task witnessed the use of simple machine learning-based solutions by teams for both the subtasks. Moreover, approaches using Term Frequency-Inverse Document Frequency (TF. IDF) feature representations and word embeddings and also machine learning algorithms including LR, RF, and Multilayer Perceptron (MLP) (Bhavukam & Kutti Padannayl, 2020; de Andrade & Gonçalves, 2020; Mishra & Mishra, 2020a, b) outperformed submissions using sophisticated transfer learning methods such as BERT. Because of the organized and competitive nature of this shared task as well as the availability of the submitted systems, this shared task could be used as a standard benchmark for research in the future.

[21] https://www.wsdm-conference.org/2020/wsdm-cup-2020.php
[22] https://lightgbm.readthedocs.io/en/latest/
[23] https://www.biendata.xyz/competition/wsdm2020/
[24] https://wosp.core.ac.uk/jcdl2020/index.html#dataset
[25] https://2020.jcdl.org/
[26] https://www.kaggle.com/c/3c-shared-task-purpose/
[27] https://www.kaggle.com/c/3c-shared-task-influence/
[28] https://www.kaggle.com/c/about/inclass

11. DISCUSSION

Early research in citation classification for identifying the reasons for citing a paper suffered several downsides. The limitations due to the size of the data sets used by such methods often resulted in low generalizability of the developed approaches. The proposed classification schemes were reported as "idiosyncratic" by White (2004) because of their domain specificity and the difficulty in application to research papers from other disciplines. The ever increasing number of scientific publications has caused severe implications related to reading all the articles manually and trying to identify their relevance. Moreover, such shortcomings resulting from manual examination of the enormous amount of documents and evaluating their importance requires remarkable domain knowledge and experience.

The advances in text and data mining techniques and the availability of infrastructures for open access full texts has steered recent research towards the development of automated methods, with promising results in this area. Researchers have developed several classification schemes with a varying number of categories to determine the citation purpose and sentiment. Another line of research, focusing on the importance of citations using a binary classifier, was also studied. In addition to instigating schemes, automated approaches also focused on testing the success of different feature sets, citation context window size, and classifiers for the effective classification of citations. Similarly, the domain also witnessed the development of several data sets for advancing research.

Despite all the advancements, there is still a lot of scope for improving the performance of the systems for citation classification. In this work, we have identified the following limitations in this field:

- **Limited size of the available data sets**—The majority of the existing domain-specific data sets contain a limited number of instances because of the difficulty of the annotation process. The recently developed larger corpora such as SciCite and the ACT data sets, which are multidomain in nature, look promising. Such data sets could enhance research in generating a cross-domain general-purpose system for citation classification.
- **Discrepancies in choosing the citation context window size**—How much information should be used for citation classification is still debated among researchers in this domain (Abu-Jbara et al., 2013; Cohan et al., 2019). Some argue that citing sentence alone is required for efficiently classifying citations, whereas others recommend the need for using additional context for classification.
- **Lack of gold standard annotated data sets for citation classification**—Another critical limitation this field has suffered is the absence of a sufficient number of large enough annotated data sets. "The success of citation classification systems depend on a small but well-defined set of citation categories" (Munkhdalai et al., 2016). The emergence of open NLP competitions such as 3C shared tasks could serve as platforms for comparing research on the same data as well as on the same classification schema. Such competitions are important in setting up a fair benchmark for evaluating methods.
- **The use of a variety of schemas makes performance comparisons difficult**—Depending on the application for which the citation classification is used, there are several classification schemas with varying complexity. As standardizing the taxonomy is difficult, comparison of the existing works is equally difficult.
- **Unbalanced nature of the available data sets**—The difficulty in obtaining annotated instances for categories, which are critical for understanding the impact produced by the citations, is yet another problem that needs to be resolved. For instance, the most used data set for citation importance classification (Valenzuela et al., 2015) has only 14% of cases

belonging to the important class. One possible reason for this is because often the authors hide their actual intentions for citing a paper in an attempt to conceal any criticism.

- **Use of objective writing style while citing a paper**—Hiding of any criticism or actual opinion in the citing sentence increases the difficulty in the detection of citation function. Use of hedging is another way of expressing uncertainty. Detection of nonexplicit reasons from the citation context is also a nontrivial problem.

The following are the potential future tasks identified by the researchers:

- **Modeling reference scope resolution**—Methods for mitigating the ambiguity caused by multiple references in the citing sentence is another area that needs more attention. Jha et al. (2017) defines reference scope resolution as methods used for identifying fragments of a sentence that are relevant to a specific target citation, given multiple references in the citing sentence. Jha et al. (2017) created a new data set for reference scope resolution with 3,500 citing sentences containing 19,591 references using AAN, as a new step towards research in this direction. CL-SciSumm[29], a shared task on scientific document summarization has a subtask for detecting the scope of the reference (Aggarwal & Sharma, 2016; Karimi, Moraes et al., 2018).
- **Use of Dynamic Citation Context**—Existing methods for citation classification use fixed context windows for extracting the linguistic features. Using fixed window size often results in either the loss of implicit citation information or the addition of noise to the citation context. NLP-based approaches for dynamically identifying the citation context still remain unexplored fully for citation classification. A recently developed data set by Lauscher, Ko et al. (2021)[30] presents the largest corpus annotated for multiple intent, which features multisentence citation context boundaries established by human annotators based on coreferences.
- **Possibility of building domain-specific models**—The domain specificity of the existing data sets resulted in research to be confined to a few individual disciplines, specifically in the Computer Science and Biomedical domains. However, scholarly publications in other fields such as Mathematics or Physics often contain equations and other mathematical symbols, which are difficult to parse. The effectiveness of domain-specific classifiers on multidomain data sets is yet to be investigated.
- **Addition of more annotations for scarce citation functions**—For mitigating the class imbalance issues of the existing data sets, use of citation function-specific annotations are recommended by researchers, to increase the number of instances in the minority classes.
- **Use of automatic methods for citation annotation**—Researchers are also considering automating the process of citation annotation with an aim to improve the problems caused by the current manual annotations. Often the complexity of the annotation schemes results in lower interannotator agreement.

Approximately 70% of the papers reviewed for citation type classification in this meta-analysis used nondeep learning-based classifiers. Such classifiers require the manual identification of features. The success of the early machine learning-based methods relied heavily on features such as dependency relations, fixed sets of cue words or phrases and other structural information which are hand-crafted and time consuming to generate. The dichotomous opinion among researchers concerning the suitability of using extended citation context for feature

[29] https://ornlcda.github.io/SDProc/sharedtasks.html#clscisumm
[30] https://github.com/allenai/multicite

extraction suggests that more research in this area is needed. Similarly, the extraction of dynamic citation contexts, which has been explored for other areas such as automatic summary generation, are yet to be studied in depth for citation function detection. Recent deep learning methods for language modeling, which are capable of capturing long-range syntactic and the semantic features from large unannotated corpora are another avenue to explore for citation classification. As authors, we look forward to the development of new general-purpose scientific models that are capable of predicting citation categories using multidomain corpora in the future.

12. CONCLUSION

Citations are critical for persuasion and are considered as a means for providing evidence or justification for authors' claims. As not all citations are equal, it is essential to understand whether the authors support or disagree with the claims made in the cited paper. This reason or author's intentions for citing a paper has long been a subject of study. In this meta-analysis, we reviewed research papers that classify citations based on their functions, polarity, and centrality. We included 60 articles in this literature review, from 1965 through to 2020. Because we gave more importance to examining the approaches that consider the discursive relations between the citing and the cited articles, 86% of the papers were from the period 2000–2020. We structured this paper based on the prototypical citation classification pipeline given in Figure 4. The Following are the important findings from this literature review.

1. The classification schemes developed for identifying citation function and polarity use low to medium to fine-grained categories. Several studies employ a hierarchical taxonomy with the lower level containing the full annotation scheme and the top level featuring more abstract classes. Citation importance classification schemes, however, use a simple binary taxonomy. The earlier data sets used for machine learning-based citation classifiers uses smaller annotated training sets, which in most cases are tagged by domain experts.
2. The nonexplicit nature of authors' intent for citing is often challenging to identify for the annotators, resulting in confusion while choosing the right category.
3. The data sources used for creating the data sets show the dominance of Computer Science (specifically Computational Linguistics) and Biomedical domains as the preferred choice. Lack of multidisciplinary data sets is a huge issue faced by this domain.
4. Several tools have been developed in the past for parsing the scientific publications, to extract the citation context and other bibliometric metadata. CRF based parsing tools such as GROBID and ParsCit continue to be used by researchers because of their effectiveness.
5. From the parsed documents, the information from citation-context is exploited for understanding the citation type. Existing research uses fixed context window sizes from one to four or more sentences surrounding the citing sentence. Researchers fall into two camps, with one group claiming the effectiveness of using a single citing sentence, whereas the other emphasizes the need for using an extended context for the successful classification of citations. This discrepancy regarding the effectiveness of using an extended context needs to be resolved and requires more investigation.
6. Classification approaches fall into three categories. The feature-based machine learning classifiers make use of contextual and/or noncontextual features, which are extracted from the citation context. Standard contextual features used by researchers are the cue words or phrases specific to the discourse structure or classes and the dependency relations, which helps capture the long-range relationship between words in the citation context. Noncontextual features such as the position of citations with respect to different sections and the frequency are vital indicators for identifying the crucial citations.

7. The recently developed deep learning methods, which do not require feeding of the handcrafted features, have shown improvement in performance when given a larger data set. However, methods using transformer architectures, such as BERT, have only been tested on simple classification schemes with three classes. The success of such models is yet to be evaluated on much broader taxonomies, which clearly distinguishes citation functions.

FUNDING INFORMATION

This research received funding from Jisc under Grant Reference: 4133, OU Scientometrics PhD Studentship, covering the contributions of Suchetha N. Kunnath and Petr Knoth.

Additional funding that contributed to the creation of the manuscript, covering the contribution of David Pride, was received from NRC, Project ID: 309594, the AI Chemist under the cooperation of IRIS.ai with The Open University, UK.

Finally, the contribution of Drahomira Herrmannova was supported by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://energy.gov/downloads/doe-public-access-plan).

AUTHOR CONTRIBUTIONS

Suchetha N. Kunnath: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing–original draft, Writing–review & editing. Drahomira Herrmannova: Formal analysis, Supervision, Validation, Writing–review & editing. David Pride: Formal analysis, Project administration, Supervision, Validation, Writing–review & editing. Petr Knoth: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing–review & editing.

COMPETING INTERESTS

The authors have no competing interests.

DATA AVAILABILITY

We did not collect any data for this research.

REFERENCES

Abu-Jbara, A., & Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 500–509). Portland, Oregon: Association for Computational Linguistics. https://aclanthology.org/P11-1051

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* (pp. 596–606). Atlanta, Georgia: Association for Computational Linguistics. https://aclanthology.org/N13-1067

Agarwal, S., Choubey, L., & Yu, H. (2010) Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings*, vol. 2010, p. 11. American Medical Informatics Association.

Aggarwal, P., & Sharma, R. (2016). Lexical and syntactic cues to identify reference scope of citance. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)* (pp. 103–112). https://aclanthology.org/W16-1512

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*. https://doi.org/10.1177/2158244019829575

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session* (pp. 81–87). Portland, Oregon: Association for Computational Linguistics. https://aclanthology.org/P11-3015

Athar, A., & Teufel, S. (2012a). Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 597–601). Montréal: Association for Computational Linguistics. https://aclanthology.org/N12-1073

Athar, A., & Teufel, S. (2012b). Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse* (pp. 18–26). Jeju Island, Korea: Association for Computational Linguistics. https://aclanthology.org/W12-4303

Bakhti, K., Niu, A., & Nyamawe, A. S. (2018). Semi-automatic annotation for citation function classification. In *2018 International Conference on Control, Artificial Intelligence, Robotics Optimization (ICCAIRO)* (pp. 43–47). https://doi.org/10.1109/ICCAIRO.2018.00016

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615–3620). Hong Kong: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Bertin, M., & Atanassova, I. (2012). Semantic enrichment of scientific publications and metadata. *D-lib Magazine*, *18*(7/8). https://doi.org/10.1045/july2012-bertin

Bertin, M., & Atanassova, I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. In *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014)* (pp. 5–12). Amsterdam.

Bertin, M., Atanassova, I., Sugimoto, C. R., & Larivière, V. (2016). The linguistic patterns and rhetorical structure of citation context: An approach using n-grams. *Scientometrics*, *109*(3), 1417–1434. https://doi.org/10.1007/s11192-016-2134-8

Bhavukam, P., & Kutti Padannayl, S. (2020). Amrita_CEN_NLP @ WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications* (pp. 71–74). Wuhan, China: Association for Computational Linguistics. https://aclanthology.org/2020.wosp-1.11

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., ... Tan, Y. F. (2008). The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*. Marrakech, Morocco: European Language Resources Association (ELRA). https://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1). https://doi.org/10.1108/00220410810844150

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science*, *36*(4), 223–229. https://doi.org/10.1002/asi.4630360402

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284–290. https://doi.org/10.1002/(SICI)1097-4571(198907)40:4<284::AID-ASI10>3.0.CO;2-Z

Chen, W., Liu, S., Bao, W., & Jiang, H. (2020). An effective approach for citation intent recognition based on BERT and lightGBM. *WSDM Cup*, Houston, Texas.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, *5*(4), 423–441. https://doi.org/10.1177/030631277500500403

Cohan, A., Ammar, W., van Zuylen, 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: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3586–3596). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1361

Constantin, A., Pettifer, S., & Voronkov, A. (2013). PDFX: Fully-automated PDF-to-XML conversion of scientific literature. In *Proceedings of the 2013 ACM Symposium on Document Engineering* (pp. 177–180). New York. Association for Computing Machinery. https://doi.org/10.1145/2494266.2494271

Councill, I., Giles, C. L., & Kan, M.-Y. (2008). ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). https://www.lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf

de Andrade, C. M. V., & Gonçalves, M. A. (2020). Combining representations for effective citation classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications* (pp. 54–58). Wuhan, China: Association for Computational Linguistics. https://aclanthology.org/2020.wosp-1.8

Di Marco, C., Kroon, F. W., & Mercer, R. E. (2006). Using hedges to classify citations in scientific articles. In *Computing attitude and affect in text: theory and applications* (pp. 247–263). Springer. https://doi.org/10.1007/1-4020-4102-0_19

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 623–631). Chiang Mai, Thailand: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1070.

Fisas, B., Ronzano, F., & Saggion, H. (2016). A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)* (pp. 3081–3088). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1492

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *The Library Quarterly*, *49*(4), 399–414. https://doi.org/10.1086/600930

Garfield, E. (1965). Can citation indexing be automated? In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation, symposium proceedings*, vol. 269, pp. 189–192. Washington.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471, PubMed: 5079701

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375. https://doi.org/10.1007/BF02019306

Garzone, M., & Mercer, R. E. (2000). Towards an automated citation classifier. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence* (pp. 337–346). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45486-1_28

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, *7*(1), 113–122. https://doi.org/10.1177/030631277700700112

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*(3), 497–518. https://doi.org/10.1016/j.pragma.2008.06.001

Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–8). https://doi.org/10.1109/JCDL.2017.7991558

Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, *116*(2), 973–996. https://doi.org/10.1007/s11192-018-2767-x

Hernández-Álvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, *22*(3), 327–349. https://doi.org/10.1017/S1351324915000388

Hernández-Álvarez, M., Gomez Soriano, J. M., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, *23*(4), 561–588. https://doi.org/10.1017/S1351324916000346

Hernández-Álvarez, M., Gómez, J. M., & Martínez-Barco, P. (2016). *Annotated corpus for citation context analysis*. https://www.semanticscholar.org/paper/Annotated-Corpus-for-Citation-Context-Analysis-Hern%C3%A1ndez-%C3%81lvarez-Soriano/c1756794d1d39be771b9f19b86bf3c64102c3476

Herrmannova, D., Patton, R. M., Knoth, P., & Stahl, C. G. (2018). Do citations and readership identify seminal publications? *Scientometrics*, *115*(1), 239–262. https://doi.org/10.1007/s11192-018-2669-y

Hou, W.-R., Li, M., & Niu, D.-K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution: Citation frequency of individual articles in other papers more fairly measures their scientific contribution than mere presence in reference lists. *BioEssays*, *33*(10), 724–727. https://doi.org/10.1002/bies.201100067, PubMed: 21826692

Ioannidis, J. P. A. (2006). Concentration of the most-cited papers in the scientific literature: Analysis of journal ecosystems. *PLOS ONE*, *1*(1), e5. https://doi.org/10.1371/journal.pone.0000005, PubMed: 17183679

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

Jochim, C., & Schütze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING 2012* (pp. 1343–1358). Mumbai, India: The COLING 2012 Organizing Committee. https://aclanthology.org/C12-1082

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391–406. https://doi.org/10.1162/tacl_a_00028

Kaplan, D., Tokunaga, T., & Teufel, S. (2016). Citation block determination using textual coherence. *Journal of Information Processing*, *24*(3), 540–553. https://doi.org/10.2197/ipsjjip.24.540

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, *16*(3), 179–184. https://doi.org/10.1002/asi.5090160305

Karimi, S., Moraes, L., Das, A., Shakery, A., & Verma, R. (2018). Citance-based retrieval and summarization using IR and machine learning. *Scientometrics*, *116*(2), 1331–1366. https://doi.org/10.1007/s11192-018-2785-8

Kunnath, S. N., Pride, D., Gyawali, B., & Knoth, P. (2020). Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications* (pp. 75–83). Wuhan, China: Association for Computational Linguistics. https://aclanthology.org/2020.wosp-1.12

Lauscher, A., Glavaš, G., Ponzetto, S. P., & Eckert, K. (2017). Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications* (pp. 24–28). https://doi.org/10.1145/3127526.3127531

Lauscher, A., Ko, B., Kuhl, B., Johnson, S., Jurgens, D., ... Lo, K. (2021). Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *arXiv preprint* arXiv:2107.00414.

Le, M.-H., Ho, T.-B., & Nakamori, Y. (2006). Detecting citation types using finite-state machines. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 265–274). Springer. https://doi.org/10.1007/11731139_32

Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards fine-grained citation function classification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 402–407). Hissar, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/R13-1052

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4969–4983). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.447

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 473–474). Springer. https://doi.org/10.1007/978-3-642-04346-8_62

Meng, R., Lu, W., Chi, Y., & Han, S. (2017). Automatic classification of citation function by new linguistic features. *iConference 2017 Proceedings*. https://doi.org/10.9776/17349

Mercer, R. E., & Di Marco, C. (2003). The importance of fine-grained cue phrases in scientific citations. In Y. Xiang & B. Chaib-draa (Eds.), *Advances in Artificial Intelligence* (pp. 550–556). Springer. https://doi.org/10.1007/3-540-44886-1_49

Mishra, S., & Mishra, S. (2020a). Scubed at 3C task A—A simple baseline for citation context purpose classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications* (pp. 59–64). Wuhan, China. Association for Computational Linguistics. https://aclanthology.org/2020.wosp-1.9

Mishra, S., & Mishra, S. (2020b). Scubed at 3C task B—A simple baseline for citation context influence classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications* (pp. 65–70). Association for Computational Linguistics. https://aclanthology.org/2020.wosp-1.10

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92. https://doi.org/10.1177/030631277500500106

Munkhdalai, T., Lalor, J. P., & Yu, H. (2016). Citation analysis with neural attention models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis* (pp. 69–77). Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-6109

Nakov, P. I., Schwartz, A. S., Hearst, M., et al. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, vol. 4, pp. 81–88. Citeseer.

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, *11*(1), 117–134. https://doi.org/10.7152/acro.v11i1.12774

Nazir, S., Asif, M., & Ahmad, S. (2020a). Important citation identification by exploiting the optimal in-text citation frequency. In *2020 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1–6). https://doi.org/10.1109/ICEET48479.2020.9048224

Nazir, S., Asif, M., Ahmad, S., Bukhari, F., Afzal, M. T., & Aljuaid, H. (2020b). Important citation identification by exploiting content and section-wise in-text citation count. *PLOS ONE*, *15*(3). https://doi.org/10.1371/journal.pone.0228885, PubMed: 32134940

Nissim, M., Abzianidze, L., Evang, K., van der Goot, R., Haagsma, H., ... Wieling, M. (2017). Last words: Sharing is caring: The future of shared tasks. *Computational Linguistics*, *43*(4), 897–904. https://doi.org/10.1162/COLI_a_00304

Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, *29*(5), 225–231. https://doi.org/10.1002/asi.4630290504

Perier-Camby, J., Bertin, M., Atanassova, I., & Armetta, F. (2019). A preliminary study to compare deep learning with rule-based approaches for citation classification. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*. Cologne.

Pham, S. B., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence* (pp. 759–771). https://doi.org/10.1007/978-3-540-24581-0_65

Prasad, A., Kaur, M., & Kan, M.-Y. (2018). Neural ParsCit: A deep learning-based reference string parser. *International Journal on Digital Libraries*, *19*(4), 323–337. https://doi.org/10.1007/s00799-018-0242-1

Pride, D., & Knoth, P. (2017a). Incidental or influential? Challenges in automatically detecting citation importance using publication full texts. In *International Conference on Theory and Practice of Digital Libraries* (pp. 572–578). Springer. https://doi.org/10.1007/978-3-319-67008-9_48

Pride, D., & Knoth, P. (2017b). Incidental or influential? A decade of using text-mining for citation function classification. In *16th International Society of Scientometrics and Informetrics Conference*. Wuhan, China.

Pride, D., & Knoth, P. (2020). An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 337–340). New York: Association for Computing Machinery. https://doi.org/10.1145/3383583.3398617

Pride, D., Knoth, P., & Harag, J. (2019). Act: An annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 329–330). IEEE. https://doi.org/10.1109/JCDL.2019.00055

Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics*, *118*(1), 21–43. https://doi.org/10.1007/s11192-018-2961-x

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, *47*(4), 919–944. https://doi.org/10.1007/s10579-012-9211-2

Radoulov, R. (2008). Exploring automatic citation classification. Master's Thesis, University of Waterloo.

Rotondi, A., Di Iorio, A., & Limpens, F. (2018). Identifying citation contexts: A review of strategies and goals. In *CLiC-it*. https://doi.org/10.4000/BOOKS.AACCADEMIA.3594

Schäfer, U., & Kasterka, U. (2010). Scientific authoring support: A tool to navigate in typed citation graphs. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids* (pp. 7–14). Los Angeles: Association for Computational Linguistics. https://aclanthology.org/W10-0402

Shotton, D. (2010). Cito, the citation typing ontology. *Journal of Biomedical Semantics*, *1*, S6. https://doi.org/10.1186/2041-1480-1-S1-S6, PubMed: 20626926

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., ... Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). New York: Association for Computing Machinery. https://doi.org/10.1145/2740908.2742839

Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, *7*(1), 97–113. https://doi.org/10.1177/030631277700700111

Su, X., Prasad, A., Kan, M.-Y., & Sugiyama, K. (2019). In *Neural Multi-Task Learning for Citation Function and Provenance* (pp. 394–395). IEEE Press. https://doi.org/10.1109/JCDL.2019.00122

Sula, C. A., & Miller, M. (2014). Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, *29*(3), 452–464. https://doi.org/10.1093/llc/fqu019

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics*, *7*(1), 39–56. https://doi.org/10.1093/applin/7.1.39

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80–87). Sydney, Australia: Association for Computational Linguistics. https://aclanthology.org/W06-1312. https://doi.org/10.3115/1654595.1654612

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). Sydney, Australia: Association for Computational Linguistics. https://aclanthology.org/W06-1613. https://doi.org/10.3115/1610075.1610091

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, *18* (4), 317–335. https://doi.org/10.1007/s10032-015-0249-8

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *Workshops at the Twenty-ninth AAAI Conference on Artificial Intelligence*.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020a). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020b). Important citation identification by exploiting the

syntactic and contextual information of citations. *Scientometrics*, *125*, 2109–2129. https://doi.org/10.1007/s11192-020-03677-1

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics*, *25*(1), 89–116. https://doi.org/10.1093/applin/25.1.89

Xu, H., Martin, E., & Mahidadia, A. (2013). Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*.

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, *335*, 195–205. https://doi.org/10.1016/j.neucom.2019.01.021

Zafar, L., Ahmed, U., & Islam, M. A. (2019). Citation context analysis using word-graph. In *2nd International Conference on Communication, Computing and Digital systems (C-CODE)* (pp. 120–125). https://doi.org/10.1109/C-CODE.2019.8680976

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427. https://doi.org/10.1002/asi.23179