A Meta-analysis of Semantic Classification of Citations

Suchetha N. Kunnath  
snk56@open.ac.uk  
Knowledge Media Institute (KMi)  
The Open University  
Milton Keynes, UK

Drahomira Herrmannova  
herrmannova@ornl.gov  
Knowledge Media Institute (KMi)  
The Open University  
Milton Keynes, UK

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 characterising 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 given on the later automated methods, which use Machine Learning and Natural Language Processing (NLP) for analysing the fine-grained linguistic features in the surrounding text of citations. The sections are organised based on the steps involved in the pipeline for citation classification. Specifically, we explore the existing classification schemes, datasets, pre-processing methods, the extraction of contextual and non-contextual 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.

KEYWORDS

Citation Classification, Citation Type, Citation Function, Citation Polarity, Citation Importance, Citation Context

1 INTRODUCTION

Citation analysis has been a subject of study for several decades with the work of Garfield (1972) being amongst the most pioneering. One of the primary motivations for studies related to bibliographic references is related to identifying methods for research assessment and evaluation (Swales, 1986). The existing methods using citation impact indicators like 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 et al., 2019). Traditional use of citation counts alone as an indicator for measuring the scientific impact of research publications, researchers, as well as research institutions has been widely criticised in the past (Kaplan, 1965; Moravcsik and Murugesan, 1975). The

San Francisco Declaration on Research Assessment (DORA) 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 publication availability (Bornmann and Daniel, 2008). How to weigh such individual factors is still unclear while 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 et al., 2017; Jurgens et al., 2018; Zhu et al., 2015). For instance, a citation that criticises a work has a different influence than a citation used as a starting point for new research (HERNÁNDEZ-ALVAREZ et al., 2017). Abu-Jbara et al. (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 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, Nigel 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 are not just interested in how many times a work is cited, but also why it is being cited (Lauscher et al., 2017). However, Nakov et al. (2004) show that there are a variety of other application areas including document summarisation, document indexing and retrieval and monitoring research trends, which can be seen as beneficiaries for 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 and Daniel, 2008; HERNÁNDEZ-ALVAREZ and GÓMEZ, 2016; Tahamtan and 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 given on the later automated methods than on the earlier manual work for citation classification. This paper is organised as follows: section 2 describes the process of citation classification, important terminologies, application and challenges in this area. Section 3 explains the methods we used for collecting research papers for this meta-analysis. The sections 4 and 5 reviews the popular classification schemes and the datasets. This is followed by examining methods used for the different steps involved in the automatic citation classification, namely pre-processing, important feature identification, classification and evaluation. Section 10 describes the open competitions in this domain.

2 CITATION CLASSIFICATION

Research publications are not stand-alone 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; Moravcsik and Murugesan, 1975; Nigel Gilbert, 1977; Oppenheim and Renn, 1978). Garfield et al. (1965) in his pioneering work identifies 15 reasons for citing a paper, of which few of them are “Paying homage to pioneers, Giving credit for related work, Identifying method, equipment etc., Providing background reading” and so forth. All these studies which developed taxonomies for characterising citations aimed at identifying the social functions that reference serves and determining how important it is to the citing author in order to shed insight about the authors citing practices (Radoulov, 2008). Earlier methods used either surveying publishing authors (Brooks, 1985; Cano, 1989) or the expertise of the analysts (Chubin and Moitra, 1975; Moravcsik and Murugesan, 1975) to decode the implicit aspects of citations from the text surrounding the reference (Sula and Müller, 2014). However, little attention was given in analysing the scientific content of the citation context.

The citation classification problem, however, from a discourse analyst point of view was later studied by (Swales, 1986; Teufel et al., 2006b; White, 2004). Here, the explicitly mentioned words or phrases surrounding the citation are analysed for interpreting 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 citations types from a range of perspectives. These include 1) understanding citation functions, which constitute the roles or purposes associated with a citation, by examining the citation context (Cohan et al., 2019; Garzone and Mercer, 2000; Jurgens et al., 2018; Teufel et al., 2006b), 2) citation

Figure 1: Steps involved in the collection of research papers for this survey
polarity or sentiment, which gives insight about the author’s disposition towards the cited document (HERNÁNDEZ-ÁLVAREZ et al., 2017; Lauscher et al., 2017), and 3) citation importance, where the citations are grouped based on how influential/important they are to the cited document (Pride and Knoth, 2017b; Valenzuela et al., 2015; Zhu et al., 2015).

The 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 non-textual 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 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 contains 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 other’s 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 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 discursive relationship between the cited and the citing text. This is followed by the subsections, challenges and applications of automatic citation classification methods.

### 2.2 Challenges

Classifying citations based on their type is not a trivial task. Firstly, the citing sentence might not always explicitly contain the necessary semantic cues enabling to determine the citation type. Secondly, 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 datasets, without explicitly mentioning the citation (Kaplan et al., 2016). Disregarding such implicit citations results in an information loss when characterising citations (Athar and Teufel, 2012b). Occasionally, authors use exaggerated praises to hide criticism, thus avoiding negative citations, and show reluctance to acknowledge using a specific method from previous research (Teufel et al., 2006a). Developing a classification scheme
that can successfully capture the broad range of citation functions too is challenging. Classification schemes often range from those rather abstract to fine-grained. While the abstract taxonomies are too general to capture all the specific information (Radoulov, 2008), the inter-annotator agreement decreases substantially in the case of the fine-grained, with the annotators experiencing difficulties in choosing between similar or overlapping categories (Agarwal et al., 2010; Hernández-Álvarez 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 et al., 2016). Additionally, most of the existing datasets for citation classification are manually annotated by domain experts, which is hugely time-consuming and therefore expensive, and also sometimes, the granularity of the fine-grained schemes is reduced due to the inter-annotator agreement decreases substantially in the case of the fine-grained, with the annotators experiencing difficulties in choosing between similar or overlapping categories (Agarwal et al., 2010; Hernández-Álvarez 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 et al., 2016). Additionally, most of the existing datasets for citation classification are manually annotated by domain experts, which is hugely time-consuming and therefore expensive, and also potentially subjective (Bakhli et al., 2018).

Progress in this field has been hampered by the lack of annotated corpora, large enough to generalise the task, and irrespective of the domain (HERNÁNDEZ-ALVAREZ and GOMEZ, 2016; Hernández-Álvarez et al., 2016; Radoulov, 2008). Non-reuse of the existing datasets, 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 and 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 et al., 2020). The domain-specific nature of existing datasets 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 datasets 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 and Schäfer, 2011; Fisas et al., 2016; Jurgens et al., 2018). Often supervised learning methods for citation classification fail to categorise citations to the minority classes, which are of more importance in this task (Dong and Schäfer, 2011).

2.3 Applications


To extract the most representative subset for citation-based summarisation, Abu-Jbara and Radev (2011) classify the initial filtered citing sentences based on the five function types: (1) Background, (2) Problem Statement, (3) Method, (4) Results, and (5) Limitations. Fisas et al. (2016) introduced a multi-layer 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 visualisation 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.

### Table 1: Reviewed papers with domain specific applications for citation classification

<table>
<thead>
<tr>
<th>Application</th>
<th>Paper</th>
<th>%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Retrieval</td>
<td>Garzone and Mercer (2000)</td>
<td>11.6%</td>
</tr>
<tr>
<td></td>
<td>Di Marco et al. (2006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Schäfer and Kasterka (2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Agarwal et al. (2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bertin and Atanassova (2012)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xu et al. (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valenzuela et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>Summarisation</td>
<td>Nanba et al. (2000)</td>
<td>6.6%</td>
</tr>
<tr>
<td></td>
<td>Abu-Jbara and Radev (2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fisas et al. (2016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>JHA et al. (2017)</td>
<td></td>
</tr>
<tr>
<td>Research Trend Detection</td>
<td>Le et al. (2006)</td>
<td>8.3%</td>
</tr>
<tr>
<td></td>
<td>JHA et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hassan et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hassan et al. (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jurgens et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>Research Evaluation</td>
<td>Moravcsik and Murugesan (1975)</td>
<td>28.3%</td>
</tr>
<tr>
<td></td>
<td>Chubin and Moitra (1975)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spiegel-Rösing (1977)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brooks (1985)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cano (1989)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Abu-Jbara et al. (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valenzuela et al. (2015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zhu et al. (2015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HERNÁNDEZ-ALVAREZ et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lauscher et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hassan et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jurgens et al. (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cohan et al. (2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quéryum and Afzal (2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yousif et al. (2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nazir et al. (2020a)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pride and Knoth (2020)</td>
<td></td>
</tr>
<tr>
<td>Venue Evolution</td>
<td>Jurgens et al. (2018)</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

* Out of total papers reviewed.

2 https://scite.ai/
for this meta-analysis. The values show that majority of the papers proposes 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, datasets and methods used by the existing systems
- Comprehend the limitations and the research gaps and
- 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
- Citation polarity
- Citation sentiment
- Citation importance
- Citation context classification
- Citation motivation
- Citation intent
- Citation purpose
- Citation behaviour and
- Citation annotation

Using these keywords, we queried the academic search engines, (1) Google Scholar³, (2) Scopus⁴, (3) ScienceDirect⁵, (4) CORE⁶ and (5) ACM Digital Library⁷. Additionally, we also searched for research papers using more generic terms like “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 5 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, datasets or methods for citation classification. Nearly 87% of the documents reviewed are from post-2000, we focused more on research corresponding to the automated approaches for citation classification. Additionally, we also review papers that discuss pre-requisite steps like scientific text extraction and pre-processing 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: (1) Manual, (2) Rule-based, (3) Machine Learning, and (4) Deep Learning, the percentage distribution of which is represented in Figure 3.

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 the subsections, citation importance and citation function schemes, both of which are utilised by the recent automated approaches.

4.1 Early Research in Citation Classification

The earliest work in citation classification is attributed to Garfield et al. (1965), who laid foundation to this domain by proposing 15 reasons why authors cite a paper. However, Garfield just defined the different categories but did not conduct in-depth research regarding the occurrence of different citation functions with respect to a paper. With the aim of determining the citation type by analysing 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 generalisability of the Moravcsik and Murugesan’s scheme by introducing a hierarchical annotation.
Figure 4: Prototypical Diagram for Citation Classification

schema featuring 6 basic classes. Using 66 articles from the Journal of 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. Since subjective opinion has more importance than factual evidence in literary research, Frost (1979) designed a classification scheme specifically for humanities. Such inter-disciplinary and intra-disciplinary variations in citation functions has been observed by researchers (Chubin and 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 7 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 et al., 2011). Moreover, most of the distinction of citations resulting from the earlier taxonomies are sociologically oriented to a greater extent and is difficult for practical application purpose (Swales, 1986; Teufel et al., 2006a). None of the schemes mentioned here makes any differentiation between self-citations; a way for manipulating citation counts and citations to other’s work (Swales, 1986). Swales (1986) raises concern as to whether it is possible to determine the intent for citing by analysing the citation context since “...the reason why an author cites as he does must remain a matter for conjecture...”. A study by Cano (1989) shows that Moravscik and Murugesan’s scheme shows the annotation of citations by authors themselves to multiple classes were paired within the expected dichotomous categories. According to the author, Moravscik and Murugesan’s citation behaviour 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 characterising 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 was conducted by Zhu et al. (2015) and Valenzuela et al. (2015). While 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 characterising both direct as well as indirect citations as incidental or important. Pride and Knoth (2017a,b) analysed 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>
<thead>
<tr>
<th>Authors</th>
<th>Classification Scheme</th>
<th>Data Source</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moravscik</td>
<td>Conceptual or Operational Use Evolutionary or Juxtapositional Organic or Perfunctory</td>
<td>Theoretical high energy physics published in Physical Review from 1968 to 1972 (inclusive)</td>
<td>30 Articles 575 references</td>
</tr>
<tr>
<td>Murugesan (1975)</td>
<td>Confirmative or Negational</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soumyo D. Moitra (1975)</td>
<td>Negative: (1) Partial, (2) Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carolyn O. Frost (1979)</td>
<td>Primary Source: (1) Supporting Factual Evidence (2) Supporting Circumstantial Evidence</td>
<td>German Literature articles from journals The Germanic Review, Euphorian, Weimarer Beitrage from years 1935, 1956, 1972</td>
<td>60 articles</td>
</tr>
<tr>
<td></td>
<td>Secondary Source: (1) Acknowledging Pioneering works (2) Indicating views on topic, (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Refer to terms/symbols (4) Support opinion, (5) Support facts, (6) Improvement of Idea,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7) Acknowledge Intellectual Indebtedness (8) Disagree with opinion, (9) Disagree with facts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10) Expressing Mixed Opinion Either Primary or Secondary: (11) Refer to further reading, (12) Provide Bibliographic Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I. Spiegel-Rösing (1977)</td>
<td>(1) Citation mentioned in Introduction/Discussion (2) Cited source is the specific point of departure for the research question</td>
<td>Social Science Citation Index (1972-1975)</td>
<td>66 articles 2309 Citations</td>
</tr>
<tr>
<td></td>
<td>(3) Cited source contains the concepts, definitions, interpretations used (4) Cited source contains data used by citing text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5) Cited source contains the data used for comparative purpose (6) Cited source contains data and material (from other disciplines than citing article)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7) Cited source contains method used (8) Cited source substantiates a statement or assumption</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(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</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Results of citing article disprove, put into question the data as interpretation of cited source (12) Results of citing article furnish a new interpretation/explanation of data of cited source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charles Oppenheim</td>
<td>(1) Historical Background (2) Description of other relevant work (3) Supplying information or data, for comparison</td>
<td>Physics and Physical Chemistry</td>
<td>23 source articles 978 citing articles (1974-1975)</td>
</tr>
<tr>
<td>Susan P. Renn (1978)</td>
<td>(4) Supplying information or data, for comparison (5) Use of theoretical equation (6) Use of methodology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7) Theory or methods not applicable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terrence A. Brooks (1985)</td>
<td>(1) Currency Scale (2) Negative Credit (3) Operational Information (4) Persuasiveness (5) Positive Credit (6) Reader Alert (7) Social Consensus</td>
<td>Multi-disciplinary</td>
<td>Papers by 26 faculties of University of Iowa</td>
</tr>
</tbody>
</table>

Table 3: Early citation function annotation schemes
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/Non-Important** and **Important/Influential**. The scheme developed by Valenzuela et al. (2015) consider 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 dataset for this task is from Valenzuela et al. (2015), using the Association for Computational Linguistics (ACL) Anthology, containing 465 citations 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 context belong to the Important or Important class for all studies. All the studies mentioned in this study used simple machine learning-based models like Support Vector Machine (SVM), Logistic Regression (LR), k-Nearest Neighbours (kNN) etc and the best performed classifier in most cases is Random Forest (RF). The most prominent predictor in all the cases is number of times a paper is cited within the citing paper (Nazir 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 which is foundational to the citing work, whilst 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 and Schäfer, 2011; Jochim and 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 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 lately 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 (1) explicitly mentioned weakness, (2) comparison or contrast, (3) agreement/usage/compatibility with the cited research and finally a (4) neutral category. (Abu-Jbara et al., 2013; JHA et al., 2017) experimented with more compressed categories containing six classes, namely, (1) Criticizing (2) Comparison, (3) Use, (4) Substantiating, (5) Basis, and (6) Neutral. The earlier schema by Moravcsik and Murugesan (1975) was later studied using automated approaches by (Dong and Schäfer, 2011; Jochim and Schütze, 2012; Meng 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 being the *Conceptual vs Operational*. Bertin and Atanassova (2012) introduced a hierarchical classification scheme with a higher level containing 5 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 Ontology³ captures the relationship between the citing and the cited articles and visualises this information using

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

Table 4: Annotation schemes and datasets used for Citation Importance classification

---

³http://purl.org/spar/cito
Semantic Web technologies (RDF, OWL etc.). A recent taxonomy introduced by scite_
classifies citation types into the classes: (1) Supporting, (2) Disrupting and (3) 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 characterising the cited articles. (Abu-Jbara et al., 2013; JHA et al., 2017; Lauscher et al., 2017; Li et al., 2013; Teufel et al., 2006a) included the categories, (1) Positive, (2) Negative, and (3) 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 citers’ motivation and the latter specifies the authors’ attitude towards the cited work. Teufel et al. (2006a,b) wrapped up the entire 12 categories as: (1) Positive - PMot, PUse, PBas, PModi, PSim, PSup, (2) Negative - Weak, CoCo- and (3) Neutral - CoCoGM, CoCoRo, CoCoXY, Neut with an aim to perform sentiment analysis over the citations.

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

5.1 Data Sources
Tables 4 and 5 show the information related to the dataset 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 task. This is largely attributed to the release of two prominent datasets for bibliographic research from ACL Anthology [10], the ACL Anthology Reference Corpus (ACL ARC) (Bird et al., 2008) and the ACL Anthology Network (AAN) corpus (Radev et al., 2013). The former consists of 10,921 articles, with full text and metadata extracted from the PDF 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 et al., 2015), is a heterogeneous graph which 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 are “…individual paragraphs immediately preceding each citation...” (Wang et al., 2020a). However, by the end of the year, 2021, Microsoft research will discontinue all MAG related services. A new Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020), which is a large English language scientific dataset, contains full text, metadata and citation links for 8.1M open access publications. This dataset is derived from sources like PubMed and arXiv.

5.2 Annotated Datasets
Table 5 shows the existing datasets for citation function classification. In an attempt to classify citations based on their rhetorical functions, Teufel et al. (2006a,b) developed a new dataset [13] using 116 conference articles and 2,829 citation instances from Computational Linguistics tagged with citation functions. Another most widely used dataset, 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 dataset was further studied extensively by (JHA et al., 2017; Lauscher et al., 2017) [14]. Jurgen et al. (2018) created a corpus with annotations for six citation functions using 585 papers from the ACL-ARC corpus [15]. The same dataset was also used by authors for experiments related to analysing the narrative structure of papers, venue evolution and modeling the evolution of the NLP field.

To address the limitations caused by the non-availability of larger annotated datasets, Cohan et al. (2019) [16] and Pride and Knoth (2020) introduced 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 multi-disciplinary dataset with 11,233 instances obtained using full text research papers from CORE. As with citation importance classification, the commonly used dataset, 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, as 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 which stated the requirement for annotating only a single “…explicitly signalled citation functions...”. The developers of the SciCite dataset used 50 test questions annotated 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, in order to improve the annotation quality. Abu-Jbara et al. (2013) sought for three different tags from the annotators: (1) Sentences relevant to citation, (2) Citation Purpose and (3) Citation Polarity. The number of annotators ranges from 2 to multiple people. Annotators in most cases are domain experts or graduate students with a background.

[10]https://www.aclweb.org/anthology/
[14]http://clair.si.umich.edu/corpora/citation_sentiment_umich.tar.gz
[16]https://github.com/allenai/scicite
<table>
<thead>
<tr>
<th>Paper</th>
<th>Classification Scheme</th>
<th>Dataset</th>
<th>Important Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garzone and Mercer 2000</td>
<td>(1) Negational - 7 classes (2) Affirmational - 5 classes (3) Assumptive - 4 classes</td>
<td>14 Journal Articles from Physics (8) and Biochemistry (6)</td>
<td>• Poor performance of classifier on unseen Physics articles (less well-structured), compared to Biochemistry articles (more well-structured).</td>
</tr>
<tr>
<td></td>
<td>(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</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanba et al. 2000</td>
<td>(1) Type B - Basis (2) Type C - Comparison or Contrast (3) Type O - Other</td>
<td>395 papers in Computational Linguistics (E-print Archive)</td>
<td>• Performance of the classifier solely depends on the cue-phrases, absence of which causes wrong prediction.</td>
</tr>
<tr>
<td>Pham and Hoffmann 2003</td>
<td>(1) Basis (2) Support (3) Limitation (4) Comparison</td>
<td>482 citation contexts and 150 unseen citation contexts</td>
<td>• Incremental knowledge acquisition using the tool, KAFTAN for citation classification</td>
</tr>
<tr>
<td>Teufel et al., 2006a</td>
<td>(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) Unfavourable Contrast/Comparison - CoCo - 1.0% (5) Contrast between 2 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%</td>
<td>116 articles and 2,829 citation instances from articles in Computational Linguistics (E-print Archive)</td>
<td>• 60% of instances belong to neutral class. • Low frequency of negative citations.</td>
</tr>
<tr>
<td>Teufel et al., 2006b</td>
<td>(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</td>
<td>811 citing areas in 9000 papers from ACM Digital Library and Science Direct</td>
<td>• Use of finite-state machines for citation type recognition does not require domain experts or knowledge about cue phrases.</td>
</tr>
<tr>
<td>Le et al., 2006</td>
<td>(1) Background/Perfunctory (2) Contemporary, (3) Contrast/Conflict (4) Evaluation, (5) Explanation, (6) Method, (7) Modality (8) Similarity/Consistency</td>
<td></td>
<td>• Model performed less on classes, Evaluation, Explanation &amp; Similarity/Consistency • Infrequent keywords not recognised by model.</td>
</tr>
<tr>
<td>Agarwal et al., 2010</td>
<td>(1) Background/Perfunctory (2) Contemporary, (3) Contrast/Conflict (4) Evaluation, (5) Explanation, (6) Method, (7) Modality (8) Similarity/Consistency</td>
<td>1,710 sentences from 43 open-access full text biomedical articles</td>
<td></td>
</tr>
<tr>
<td>Paper</td>
<td>Classification Scheme</td>
<td>Dataset</td>
<td>Important Findings</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Shotton, 2010</td>
<td>Factual: (1) cites, (2) citesAsAuthority (3) isCitedBy, (4) citesAsMetadataDocument (5) citesAsSourceDocument, (6) citesForInformation (7) obtainsBackgroundFrom (8) sharesAuthorsWith (9) usesDataFrom, (10) usesMethodIn</td>
<td>Ontology developed for Biomedical Articles</td>
<td>OWL-based tool, CiTO for characterising the nature of citations.</td>
</tr>
<tr>
<td>Dong and Schäfer 2011</td>
<td>Rhetorical - Positive: (1) confirms, (2) credits, (3) extends, (5) obtainsSupportFrom, (6) supports</td>
<td>1768 instances &amp; 122 papers from ACL Anthology (2007 and 2008)</td>
<td>Use of Ensemble-style self-training reduces the manual annotation work</td>
</tr>
<tr>
<td>Jochim and Schütze 2012</td>
<td>Rhetorical - Negative: (1) corrects, (2) critiques, (3) disagreesWith, (4) qualifies, (5) refutes</td>
<td>84 papers and 2008 citation from papers in 2004 ACL Proceedings (ARC)</td>
<td>Annotation of 4 facets instead of a single label.</td>
</tr>
<tr>
<td>Abu-Jbara et al. 2013</td>
<td>Purpose: (1) Criticizing - 14.7%, (2) Comparison - 8.5%, (3) Use - 17.7%, (4) Substantiating - 7%, (5) Basis - 7.8%, (6) Neutral - 47%</td>
<td>3,271 instances from 30 papers in ACL Anthology Network (AAN)</td>
<td>47% of citations belong to the class Neutral.</td>
</tr>
<tr>
<td>Xu et al. 2013</td>
<td>Polarity: (1) Positive - 30%, (2) Negative - 12%, (3) Neutral - 58%</td>
<td>ACL Anthology Network corpus (AAN)</td>
<td>Self-citations are skewed to the class Functional. Authors citing more has more functional citations.</td>
</tr>
<tr>
<td>Li et al., 2013</td>
<td>(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) Co-citation - 33.3%, (11) Neutral, (12) Negative</td>
<td>91 Biomedical articles and 6,355 citation instances from Biomedical articles (PubMed)</td>
<td>Coarse-grained sentiment classification performs only slightly better than fine-grained citation function classification.</td>
</tr>
<tr>
<td>HERNÁNDEZ-ALVAREZ et al. 2016</td>
<td>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%</td>
<td>2,092 citations in 85 papers from ACL Anthology Network (AAN)</td>
<td>Classes Acknowledge and Useful dominate the data distribution for purpose classification. Neutral class has more than 50% of instances.</td>
</tr>
<tr>
<td>Paper</td>
<td>Classification Scheme</td>
<td>Dataset</td>
<td>Important Findings</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Munkhdalai et al., 2016</td>
<td>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%</td>
<td>Data 1 - 3,422 (Function), 3,624 (Polarity) citations Data 2 - 4,423 (Function), 4,423 (Polarity) citations from 2,500 randomly selected PubMed Central articles</td>
<td>• Majority of citations annotated as results and findings. • Bias of citations towards positive statements.</td>
</tr>
<tr>
<td>Fisas et al., 2016</td>
<td>(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</td>
<td>10,780 sentences from 40 papers in Computer Graphics</td>
<td>• A multi-layered corpus with sentences annotated for (1) Citation purpose, (2) features to detect scientific discourse and (3) Relevance for summary.</td>
</tr>
<tr>
<td>JHA et al., 2017</td>
<td>Same as Abu-Jbara et al., 2013</td>
<td>3500 citations in 30 papers from ACL Anthology Network (AAN)</td>
<td>• Developed datasets for reference scope detection and citation context detection. • Comprehensive study aimed at applications of citation classification.</td>
</tr>
<tr>
<td>Lauscher et al., 2017</td>
<td>Same as Abu-Jbara et al., 2013</td>
<td>Datasets from (Abu-Jbara et al., 2013) and (JHA et al., 2017)</td>
<td>• Heavy skewness of dataset towards less informative classes for both schemes. • Use of Domain-specific embeddings does not enhance results.</td>
</tr>
<tr>
<td>Jurgens et al., 2018</td>
<td>(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%</td>
<td>1,969 instances from ACL-Anthology Reference Corpus (ACL-ARC)</td>
<td>• Majority of instances belong to class Background. • Error analysis shows the importance of citation context identification for result improvement.</td>
</tr>
<tr>
<td>Su et al., 2019</td>
<td>(1) Weakness - 2.2% (2) Compare and Contrast - 6.6% (3) Positive - 20.6% (4) Neutral - 70.6%</td>
<td>ACL-ARC Computational Linguistics</td>
<td>• Highly skewed dataset with majority of instances belonging to Neutral class. • Use of Multi-task learning for citation function and provenance detection.</td>
</tr>
<tr>
<td>Cohan et al., 2019</td>
<td>(1) Background - 58% (2) Method - 29% (3) Result Comparison - 13%</td>
<td>6,627 papers and 11,020 instances from Semantic Scholar, (Computer Science &amp; Medicine)</td>
<td>• Introduction of new dataset known as SciCite. • The best state-of-the-art macro-fscore obtained using BiLSTM attention with ELMo vector &amp; structural scaffolds</td>
</tr>
<tr>
<td>Pride et al., 2019</td>
<td>(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%</td>
<td>Multidisciplinary dataset of 11,233 instances from CORE</td>
<td>• Largest multi-disciplinary author annotated dataset.</td>
</tr>
</tbody>
</table>

Table 5: Citation Purpose and Polarity Classification Schemes
in the subject (Bakhti et al., 2018; Fisas et al., 2016; HERNÁNDEZ-ALVAREZ 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, specialised tools are used in certain cases. For example, Jurgens et al. (2018) employed Brat rapid annotation tool17 and two NLP experts for doubly annotating citations. (Fisas et al., 2016; Jochim and Schütze, 2012; Pride et al., 2019; Radoulov, 2008; Teufel et al., 2006a) developed web-based annotation tools for simplifying the task. To compute the agreement between the annotators, measures like Kappa coefficient (Abu-Jbara et al., 2013; Agarwal et al., 2010; Dong and Schäfer, 2011; Teufel et al., 2006a), Cohen’s Kappa coefficient, Krippendorff coefficient (HERNÁNDEZ-ALVAREZ et al., 2017) and other confidence scores (Cohan et al., 2019) are utilised. Citation annotations by independent annotators is a difficult task because often authors do not always state their intentions for citing explicitly (Nigel 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 are also been utilized for tagging citation labels (Cohan et al., 2019; Munkhdalai et al., 2016; Pride et al., 2019; Su et al., 2019).

6 PRE-PROCESSING

Text pre-processing is typically applied prior to undertaking citation function and importance classification. The process typically involves extraction of 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) like ParsCit (Councill et al., 2008), GROBID (Lopez, 2009), CERMINE (Tkaczyk et al., 2015) and Science Parse18 aim at converting the plain text or PDFs to a more semi-structured format like 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 thirteen 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 processing directly the pdf files. Other tools for extracting the in-text citations are PDFX (Constantin et al., 2013), Crossref pfdextract19, and Neural ParsCit (Prasad et al., 2018), where the former two are rule-based and the later employs Long Short Term Memory (LSTM) neural networks.

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; 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 of sentences with indirect and implicit references to the cited paper, surrounding the citing sentence are also studied for improved citation classifier performance (Athar and Teufel, 2012b; HERNÁNDEZ-ALVAREZ and GOMEZ, 2016). Rotondi et al. (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 pre-requisite 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 the 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: (1) Fixed number of characters – use of 200 characters by ParsCit20 (Jurgens et al., 2018), (2) Citing sentence – (Bertin et al., 2016; Cohan et al., 2019; Garzone and Mercer, 2000; Hassan et al., 2018; Pride et al., 2019; Sula and Miller, 2014; Valenzuela et al., 2015) and (3) 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 and Teufel, 2012a; HERNÁNDEZ-ALVAREZ 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 and Teufel, 2012b).

The usability of extended context for performance improvement has always encountered the following two concerns among the researchers, (1) the introduction of noise while incorporating additional context (Cohan et al., 2019) and (2) the loss of information in the case of using just the citing sentence for citation classification (Athar and Teufel, 2012b). Abu-Jbara et al. (2013) use sequence labeling technique for identifying the citation context. The authors

17https://brat.nlplab.org/
18https://github.com/allenai/science-parse
19https://github.com/Crossref/pfdextract
20https://parscit.comp.nus.edu.sg/
found that a window size of 4 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; 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 and Teufel, 2012b; Di Marco et al., 2006).

6.3 Mitigating Dataset Skewness
A major problem concerning the citation classifiers’ performance issues is attributed to the highly skewed nature of the classes. Several datasets report a higher number of instances for the non-important citation types like Background or Neutral and a relatively lower number of cases for more important categories like 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 dataset. The use of category-specific annotations for increasing the number of instances in the rare classes is also employed to mitigate the class-imbalance problem (Jürgens et al., 2018; Li et al., 2013; Zafar et al., 2019). Jürgens et al., 2018; Nazir et al., 2020b; Qayyum and Afzal, 2019) applied SMOTE to create synthetic instances to tackle the skewness in the dataset. 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. Analysing the SciCite dataset, Pride and Knoth (2020) found out that authors used oversampling technique on the under-represented Methods class in the dataset.

7 FEATURES FOR CITATION CLASSIFICATION
The automatic citation classification based on machine learning methods make 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 analysed 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 categorised at a higher level as Syntactic and Semantic, according to how and why the citations are described in the text. The later is further classified as (1) Textual-based, (2) Similarity-based and (3) 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 and Schäfer, 2011; Jochim and Schütze, 2012; Li et al., 2013; Meng et al., 2017). (Bertin and Atanassova, 2014; Bertin et al., 2016) emphasises the importance of verbs in understanding the nature of the relation between the citing and the cited articles. Dong and Schäfer (2011) reported best results for ensemble classifier using the syntactic POS tag features specific to each class. The application of syntactic features alone resulted in the performance improvement as compared to the baseline model for (Jochim and Schütze, 2012; 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 meta-discourse or cue words/phrases for automatic citation classification has been extensively studied in the past (Dong and Schäfer, 2011; Jürgens et al., 2018; Mercer and 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, 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 or phrases like “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). Jürgens 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. (Hasan et al., 2017; 2018; Pride and Knoth, 2017a) operationalise 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), 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 like 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 et al., 2019; Cohan et al., 2019; Perier-Camby 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; 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; Jochim and
Schütze, 2012). Jochim and Schütze (2012) also used general purpose polarity, 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 Non-Contextual Features
We categorise any extra-textual 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 and 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 clause. (Bertin and Atanassova, 2014; 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); 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 like 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 sections 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, co-authored the cited paper (Abu-Jbara et al., 2013; JHA et al., 2017; Jochim and Schütze, 2012; Jurgens et al., 2018; Teufel et al., 2006b; Zhu et al., 2015). Xu et al. (2013) identified that self-citations are prominent to the class functional, which suggests that authors new research is build on their previous works. Network based features like 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 the prior approaches, attempts were made to automate the process. The following sections discuss the existing automatic citation classification methods.

8.1 Rule-based Methods
Garzón and Mercer (2000) introduced the first automated rule-based method, where the authors categorised 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; 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 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 was employed by several studies, which applied SVM (Bakhtii et al., 2018; Hassan et al., 2017; HERNÁNDEZ-ALVAREZ 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 and Knoth, 2017a; Valenzuela et al., 2015), Naive Bayes (NB) (Abu-Jbara et al., 2013; Agarwal et al., 2010; Dong and Schäfer, 2011; Sula and Miller, 2014), Maximum Entropy (MaxEnt) (Jochim and Schütze, 2012) and so forth for training the model.

Unlike the usual supervised learning approaches, Dong and Schäfer (2011) used a semi-supervised ensemble learning model in an attempt to reduce the manual annotation of the training data. The authors used a self-training algorithm to extend the training dataset by using the predictions from the algorithm as labels for the unlabeled dataset. 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 semi-supervised method, were the authors first trained the model using NB on a small dataset and later expanded the training set using Expectation-Maximisation (EM) algorithm.

A major short-coming 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 rely 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 datasets 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 still can 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

Downloaded from http://direct.mit.edu/qss/article-pdf/doi/10.1162/qss_a_00159/1966013/qss_a_00159.pdf by guest on 21 November 2021
<table>
<thead>
<tr>
<th>Papers</th>
<th>Features Used</th>
<th>Contextual</th>
<th>Non-Contextual</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teufel et al. 2006b</td>
<td>• Verb Tense, • Voice, • Modality, • Location within (1)Article, (2)Paragraph, (3)Section</td>
<td>• Cue phrases</td>
<td>• Location within (1)Article, (2)Paragraph, (3)Section</td>
<td>• Self-citation</td>
</tr>
<tr>
<td>Dong and Schäfer 2011</td>
<td>• POS Tags, • Location within section, • Number of (1)Adjectives, (2)Adverbs, (3)Pronouns, (4)Modals &amp; (5)Cardinals, (6)Negation phrases, (7)Valance shifters</td>
<td>• Cue Words specific to classes</td>
<td>• Popularity, Density, • Avg Density</td>
<td></td>
</tr>
<tr>
<td>Athar 2011</td>
<td>• POS Tags, • Dependency Relations, • Dependency root node, • Main verb, • First person POS, • Third person POS, • Comparative/superlative POS, • Has &quot;but&quot;, • Has &quot;cf&quot;</td>
<td>• n-grams (n=1-3)</td>
<td>• General polarity lexicon, • General positive lexicon, • General negative lexicon</td>
<td></td>
</tr>
<tr>
<td>Jochim and Schütze 2012</td>
<td>• Dependency Relations, • POS Tag patterns, • Citation is a constituent, • 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 &quot;but&quot;, • Has &quot;cf&quot;</td>
<td>• Location within section</td>
<td>• Popularity, Density, • Avg Density</td>
<td>• Self-citation, • Has resource, • Has tool</td>
</tr>
<tr>
<td>Xu et al. 2013</td>
<td>• Whether citations used in parenthesis, • n-grams (n=1,2,3)</td>
<td>• Location within paper</td>
<td>• Number of citation anchors within sentence, • Self-citation</td>
<td></td>
</tr>
<tr>
<td>Li et al. 2013</td>
<td>• POS Tags, • Dependency Relations, • 3rd person pronoun, • Cue word/phrase</td>
<td>• n-grams (n=2-3)</td>
<td>• Paper relationships, • Centrality measures, • Self-citations</td>
<td></td>
</tr>
<tr>
<td>Abu-Jbara et al., 2013, JHA et al., 2017</td>
<td>• Closest Verb/Adjective/Adverb, • Contains 1st/3rd person pronoun, • Contrary Expressions, • Citation prototypicality, • Whether used in nominative/parenthetical form, • Whether preceded by (1)Pascal-cased word, (2)All-capital case word</td>
<td>• Section</td>
<td>• Reference Count, • Is Separate, • Self Citation</td>
<td></td>
</tr>
<tr>
<td>Bakhti et al. 2018</td>
<td>• Verb Tense, • Lengths of sentence and clause, • Used with Parenthesis, • Bootstrapped and Custom function patterns, • Citation prototypicality, • Whether used in nominative/parenthetical form, • Whether preceded by (1)Pascal-cased word, (2)All-capital case word</td>
<td>• Extended Cue phrases (Teufel et al., 2006b), • Citation context topics</td>
<td>• Location within (1)Paper, (2)Section, (3)Subsection, (4)Sentence, (5)Clause, • Fraction of bibliography used by reference, • Citation in (1)Subsection, (2)Sentence, (3)Clause, • Common Citations count</td>
<td>• Self-citation, • Year difference in publication dates, • Citing paper’s venue, • Reference’s venue, • Reference’s citation count &amp; PageRank, • Reference’s Hub &amp; Authority scores &amp; Network Centrality</td>
</tr>
<tr>
<td>Jurgens et al. 2018</td>
<td>• Verb Tense, • Lengths of sentence and clause, • Used with Parenthesis, • Bootstrapped and Custom function patterns, • Citation prototypicality, • Whether used in nominative/parenthetical form, • Whether preceded by (1)Pascal-cased word, (2)All-capital case word</td>
<td>• Extended Cue phrases (Teufel et al., 2006b), • Citation context topics</td>
<td>• Location within (1)Paper, (2)Section, (3)Subsection, (4)Sentence, (5)Clause, • Fraction of bibliography used by reference, • Citation in (1)Subsection, (2)Sentence, (3)Clause, • Common Citations count</td>
<td>• Self-citation, • Year difference in publication dates, • Citing paper’s venue, • Reference’s venue, • Reference’s citation count &amp; PageRank, • Reference’s Hub &amp; Authority scores &amp; Network Centrality</td>
</tr>
</tbody>
</table>

**Table 6: Features used for citation function classification**
<table>
<thead>
<tr>
<th>Papers</th>
<th>Features Used</th>
<th>Syntactic</th>
<th>Contextual</th>
<th>Semantic</th>
<th>Positional-based</th>
<th>Frequency-based</th>
<th>Other</th>
</tr>
</thead>
</table>
| Zhu et al. 2015 | • Explicit reference of cited author  
(1) Whether citations 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  
5. * of 1. Strong &  
2. Active words  
6. Word-net features  
7. General Inquirer features | • Similarity between Cited Title and  
1. Title, 2. Abstract  
3. Introduction,  
4. Conclusion &  
5. Core sections  
6. Similarity between citation context and  
1. Title, 2. Abstract  
3. Introduction  
4. Conclusion  
7. * of positive words in citation-context  
8. Emotion Lexicon for detecting  
1. Sentiment and  
2. Emotive words  
9. Citation counts in  
1. Entire Paper  
2. Introduction  
3. Related Work  
4. Core Sections  
5. # of sections where reference appears  
6. Global citations | • Whether citations appear at the  
1. Beginning or  
2. End of the sentence  
3. Position of citing sentence based-on  
1. Mean, 2. Standard variance, 3. First  
4. Last | • Citation appears in  
1. Table or caption  
2. Citation count for reference  
3. # of citations from citing to cited paper | • Self-citations  
• Publication year  
• Author overlap  
• PageRank  
• Field of cited paper  
• Author Overlap  
• Bibliographically coupled references  
| Valenzuela et al. 2015 | • Citation considered helpful based on cue phrases | • Similarity between abstracts | • Citation appears in table or caption | • Direct citations  
• # of references  
• # of paper citations / all citations  
• # of total citing papers after transitive closure | | |  
| Hassan et al. 2017  
Hassan et al. 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 cited to cited paper | • Citation frequency  
• Citation frequency  
• Citation frequency  
• Time  
• Distance  
• Author Overlap  
• Total citation length  
• Average citation length  
• Maximum citation length | |  
| Qayyum and Afzal 2019 | • Cue words  
1. n-gram similarity and dissimilarity between titles  
(n=1-3)  
2. Ratio of keywords similarity to dissimilarity between pairs  
3. Abstract similarity  
4. Similarity score  
5. Textual Similarity  
6. Similarity score  |
| Nazir et al., 2020a  
2020b | • Similarity score | | | | | |  
| Nazir et al. 2020b | • Section-wise weights for in-text citations | | | | | |  
| Wang et al. 2020b | | | | | | |  

Table 7: Features used for citation importance classification
identify features automatically, removing the pain for defining hand-crafted 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 dataset. The authors emphasise the need for larger datasets 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 multi-task citation function and sentiment analysis was proposed by Yousif et al. (2019). A multi-task 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 bi-directional LSTM and attention mechanism, along with the ELMo vectors. A new transformer based model using BERT architecture, trained on 1.14 Million scientific publications, 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 8.1M open access full text scholarly publications.

9 EVALUATION METHODS

Table 8 shows the evaluation metric and the scores obtained on the most common datasets for citation classification. The frequently used evaluation method is macro averaged F-score because of the highly skewed nature of the datasets 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. Under represented categories of the fine-grained schemes reduces 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). Since all evaluation scores mentioned in Table 8 are obtained under different settings of annotation schemes, classifiers and datasets, a comparison of methods is nearly impossible.

10 SHARED TASKS

Recent years have witnessed the increasing popularity of shared tasks, usually organised as part of conferences or workshops. The intention here is to allow research improvements in the under-researched or under-resourced areas of NLP, thus making possible the comparison of competing systems in such competitions (Nissim et al., 2017). Although research into citation function has made considerable progress since late 1970’s, 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 organised in the year 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, organised by Microsoft research as part of the WSDM Cup 202021 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 top 3 papers based on their relevance from a database. Using the described text as query, the participating teams should be able to retrieve the candidate papers from a pool of over 800k 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 et al., 2020). This shared task was hosted on the data science competition hosting platform, Biendata23.

10.2 3C Citation Context Classification Task

The 3C citation context classification task (Kunnath et al., 2020) organised by The Open University, UK as part of the workshop, WOSP 202024 and collocated with JCDL 202025, was the first shared task featuring the classification of citations based on its purpose and influence. This task utilised a portion (3,000 training instances) of the new multi-disciplinary ACT dataset (Pride et al., 2019), the largest dataset annotated by authors themselves. The 3C shared task was organised as two subtasks; (1) Subtask A - Citation Context Classification based on purpose26, a multi-class classification problem based on the citation functions and (2) Subtask B - Citation Context Classification based on influence27, a binary task focusing on the citation importance classification. Both these subtasks were hosted as separate competitions using the Kaggle InClass competitions28.

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. A total of four teams participated in this shared task, of which, three teams competed in both the tasks. All systems submitted were evaluated using macro averaged F-score on a test set of 1,000 instances. Despite the recent advancements in the 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, Multi-Layer Perceptron (MLP) (Bhavukam and Kutti Padannay, 2020; de Andrade and Gonçalves, 2020; Mishra and Mishra, 2020a;b) outperformed submissions using sophisticated transfer learning methods such as BERT. Because of the organised and competitive nature of this shared task as well as the availability of the submitted

---

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
systems, this shared task could be used as a standard benchmark for research in the future.

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 datasets used by such methods often resulted in low generalisability 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 advancements 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 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 datasets 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 datasets** – The majority of the existing domain-specific datasets contain a limited number of instances because of the difficulty in the annotation process. The recently developed larger corpora like SciCite and the ACT datasets, which are multi-domain in nature, look promising. Such datasets 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 a debate 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, while others recommend the need for using additional context for classification.

- **Lack of gold standard annotated datasets for citation classification** – Another critical limitation this field has suffered is the absence of a sufficient number of large enough annotated datasets. “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 like 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. Since standardising the taxonomy is difficult, comparison of the existing works is equally difficult.

- **Unbalanced nature of the available datasets** – 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 dataset 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 non-explicit reasons from the citation context is also a non-trivial problem.

The following are the potential future works as 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 dataset for reference scope resolution with 3,500 citing sentences containing a total of 19,591 references using AAN, as a new step towards research in this direction. CL-SciSumm29, a shared task on scientific document summarisation has a subtask for detecting the scope of the reference (Aggarwal and Sharma, 2016; Karimi 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

remain still unexplored fully for citation classification. A recently developed dataset by Lauscher et al. (2021)30 presents the largest corpus annotated for multiple intent, which features multi-sentence citation context boundaries established by human annotators based on co-references.

- **Possibility of building domain-specific models** – The domain specificity of the existing datasets resulted in research to be confined to a few individual disciplines, specifically in Computer Science and Biomedical domain. However, scholarly publications in other fields like Mathematics or Physics, often contain equations and other mathematical symbols which are difficult to parse. The effectiveness of domain-specific classifiers on multi-domain datasets is yet to be investigated.

- **Addition of more annotations for scarce citation functions** – For mitigating the class imbalance issues of the existing datasets, 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 inter-annotator agreement.

Approximately 70% of the papers reviewed for citation type classification in this meta-analysis used non-deep learning-based classifiers. Such classifiers require the manual identification of features. The success of the early machine learning-based methods relied heavily on features like dependency relations, fixed set 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 extraction suggests more research in this area is needed. Similarly, the extraction of dynamic citation contexts which has been explored for other areas like 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 multi-domain 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. Since 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. Since we are gave more importance to

29https://ornlclda.github.io/SDProc/sharedtasks.html#clscisumm

30https://github.com/allenai/multicite
examining the approaches that considers the discursive relations between the citing and the cited articles. 86% percentage of the papers were from the period 2000 - 2020. We structured this paper based on the prototypical citation classification pipeline given in Figure 4. 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 hierarchical taxonomy with lower level containing the full annotation scheme and the top-level featuring more abstract classes. Citation importance classification schemes, however, use simple binary taxonomy. The earlier datasets used for machine learning-based citation classifiers uses smaller annotated training sets which in most cases are tagged by domain experts.

(2) The non-explicit nature of authors intent for citing is often challenging to identify for the annotators, resulting in confusions while choosing the right category.

(3) The data sources used for creating the datasets shows the dominance of Computer Science (specifically Computational Linguistics) and biomedical domains as the preferred choice. Lack of multi-disciplinary datasets 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 like 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 1 to 4 or more sentences, surrounding the citing sentence. Researchers fall into two camps, with one group claiming the effectiveness of using a single citing sentence, whilst the other emphasise 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 the contextual and/or non-contextual 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. Non-contextual features like 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 hand-crafted features have shown improvement in performance when given a larger dataset. 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. Kummath 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 United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States 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 (http://energy.gov/downloads/doe-public-access-plan).

AUTHOR CONTRIBUTIONS
Suchetha N. Kummath: 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, Validation, Writing–review & editing.

COMPETING INTERESTS
The authors have no competing interests.

DATA AVAILABILITY
We did not collect any data for this research.

REFERENCES


