Journal of Informetrics 19 (2025) 101608

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research Paper

Deep learning meets bibliometrics: A survey of citation function classification ☆

Yang Zhang [a,b,*], Yufei Wang [a], Quan Z. Sheng [a], Lina Yao [e,a], Haihua Chen [b], Kai Wang [d,**], Adnan Mahmood [a], Wei Emma Zhang [f], Munazza Zaib [a], Subhash Sagar [a], Rongying Zhao [c]

[a] Macquarie University, Sydney, 2113, NSW, Australia
[b] University of North Texas, Denton, 76201, TX, USA
[c] Wuhan University, Wuhan, 430000, Hubei, China
[d] Nanyang Technological University, 639798, Singapore
[e] CSIRO's Data61, Sydney, 2015, NSW, Australia
[f] The University of Adelaide, Adelaide, 5005, NT, Australia

ARTICLE INFO

Keywords:
Citation function
Deep learning
Bibliometrics
Pretrained language model

ABSTRACT

With the advent and progression of Natural Language Processing (NLP) methodologies, the domain of automatic citation function classification has gained popularity and considerable research efforts have been contributed to this task. Automatic citation function classification has a joint computational linguistic and bibliometrics background. However, due to the different expertise in both fields, there is rarely a comprehensive and unified analysis of this task. We provide a detailed and nuanced examination analysis of the evolution of citation function classification task from the dimensions of citation function annotation schemes, widely employed benchmarks, and computational models. We first present the origins and the development of the citation function classification task. From the perspective of multi-disciplinary integration, we then discuss how bibliometrics and NLP can be better combined to contribute to the citation function classification task. Finally, based on the deficiencies that we have found in the task, we suggest some promising prospects in both bibliometrics and NLP to be investigated.

1. Introduction

Citations provide a way of validating the contributions of a specific scientific paper and connecting scientific papers in a unified citation network. They are an essential tool for understanding the context and motivations of a scientific text, and for tracing the development and evolution of ideas and research within a particular field (Shadish et al., 1995; Zhang et al., 2021; Aljohani et al., 2023). For many years, the most widely used metric to evaluate the scientific impact has been the citation counts (Hassan et al., 2020; Qian et al., 2020). Many different computational indicators based on citation counts have been developed to assess the performance of scholars, individuals, and institutions (Safer & Tang, 2009; Jung & Segev, 2013; Berrebbi et al., 2022; Hu et al., 2023; Zhang et al.,

☆ This study is supported by ARC (Australian Research Council) (Grant Number: LP190100140 and DP230100233).
* Corresponding author at: Macquarie University, Sydney, 2113, NSW, Australia.
** Corresponding author.
E-mail addresses: Yang.zhang@mq.edu.au (Y. Zhang), kai_wang@ntu.edu.sg (K. Wang).

https://doi.org/10.1016/j.joi.2024.101608
Received 17 March 2024; Received in revised form 22 October 2024; Accepted 11 November 2024

Available online 14 November 2024

Table 1
Previous surveys related to citation function classification. Biblio.–Bibliometrics; C.S.–Computer Science.

Paper Year	Research Method	Domains	Research Focus	Deep Learning	Pretrained Language Model	Comparison of Annotation Scheme/Datasets	Comparison of Models
2014	Non-systematic	Biblio.	The paper provides an insightful survey of citation content analysis, showcasing current methodologies and applications.	NO	NO	NO	NO
2016	Non-systematic	C.S.	The paper summarizes recent NLP-driven citation analysis research, emphasizing significant experiments and practical examples.	YES	NO	NO	NO
2017	Non-systematic	C.S.	The paper condenses research on identifying and classifying citation contexts, exploring latest techniques and data repositories used.	YES	NO	YES	NO
2021	Systematic	Biblio.	The paper employs a meta-synthesis approach to create a new classification for citation motivations.	YES	YES	NO	NO
2021	Systematic	C.S.	The paper examines publications using NLP and machine learning for citation related tasks	YES	YES	YES	NO
2023b	Non-systematic	C.S.	The paper explores the symbiotic relationship between citations and large language models, highlighting their joint advancements and impacts.	YES	YES	NO	NO

2024). However, citations are used in scientific papers for many reasons (Small, 1982; Siddharthan & Teufel, 2007; Zhu et al., 2015; Jung & Segev, 2013). For example, some citations are used to describe the background information about the research topics, whereas others may provide motivation or technical knowledge. Knowing these citation functions can facilitate further research, including but not limited to, identifying meaningful citations (Valenzuela et al., 2015), improving the efficiency of literature review (Lyu et al., 2021), and creating better citation-based indicator (Teufel et al., 2006b).

In 1965, Garfield, the pioneer of bibliometrics, identified fifteen different reasons why a paper might be cited (Garfield, 1965). This is also one of the earliest efforts to consider the different types of citations. However, in the past century, scholars primarily used manual counting methods to differentiate between different types of citations (Weinstock, 1971; Moravcsik & Murugesan, 1975). With the proliferation of scientific publications on the Internet, the automated classification of citations has gained significant importance. With the advancements in text mining and NLP techniques, computer science researchers are now able to process large amounts of text, thereby making it possible to automate citation function classification. Teufel et al. (2006b) proposed an automated classification model based on supervised learning. Following this, the citation function classification task becomes more and more popular among computer science and data science scholars, particularly when the deep learning model becomes more powerful to tackle the textual information (Abu-Jbara et al., 2013; Jurgens et al., 2018; Cohan et al., 2019; Zhang et al., 2021; Du et al., 2023).

Existing surveys on Citation Function Classification from the bibliometrics domain (Ding et al., 2014; Lyu et al., 2021) introduce a comprehensive set of citation function labels and methods but overlook computational feasibility. In contrast, NLP-focused surveys by Hernández-Alvarez and Gomez (2016), Jha et al. (2017), and Iqbal et al. (2021) emphasize text analysis tasks and advanced algorithms, often integrating citation function classification within broader citation content analysis frameworks. Unlike previous works, this survey evaluates the citation function classification task from both bibliometrics and modern NLP perspectives. We systematically review research efforts in citation function classification, focusing on annotation schemes, popular benchmarks, and state-of-the-art computational models from the past decades.

We come to the following conclusions: 1) the annotation schemes are becoming increasingly simple, 2) the scale of datasets is becoming increasingly large, and 3) the computational models used for citation function classification are becoming more and more computationally expensive and complicated. Additionally, we identify the current research challenges and summarize the future directions in citation function classification, which are poised to benefit both the bibliometrics and computer science research communities significantly. Our contributions are three folds:

- We introduce the evaluation for the task of *Citation Function Classification* from bibliometrics to modern NLP as well as the contributions from both fields.
- We systematically and comprehensively review the annotation scheme, benchmarks, and computational models for the task of *Citation Function Classification*.
- We summarize the major problems of existing research efforts in *Citation Function Classification* and point out several promising research directions.

2. Literature review

There are some existing surveys covering the *Citation Function Classification* task. From the bibliometrics domain, Ding et al. (2014) introduce the concept of citation function classification task as well as its future development. Lyu et al. (2021), on the other hand, focus on introducing a diverse and comprehensive set of citation function labels based on the previous research efforts. However, both of them do not consider computational feasibility. In contrast, from the NLP domain, Hernández-Alvarez and Gomez (2016) focuses on describing all text analysis tasks associated with citations, wherein citation function classification is just one of the tasks. Jha

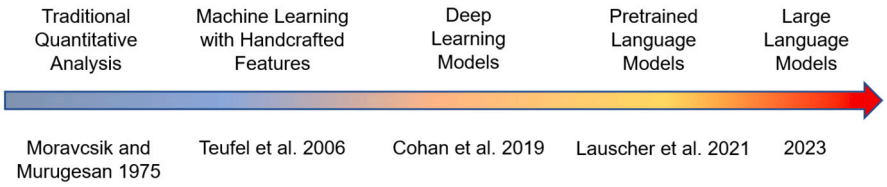

Fig. 1. Representative works of different technologies in different periods of citation function classification.

et al. (2017) and Iqbal et al. (2021) simply list the advanced machine learning and NLP algorithms employed in citation function classification tasks. We also list some other highly cited surveys that are relevant to citation function classification in Table 1.

We further compare some highly cited surveys related to citation function classification published in recent ten years, including their research method and review focus. Most of them are not only focused on citation function classification but also include other citation content analysis. However, prior reviews in the field of bibliometrics have primarily focused on providing an introduction to citation content analysis and discussing the concept of the citation function classification task in a general sense. Additionally, researchers in the bibliometrics domain investigate different citation functions and further provide more logical citation functions to the research community. Researchers from computer science always have a different point of view when they draft the literature review related to citation function classification as they mostly focus on introducing the citation function classification task in a computational perspective (Hernández-Alvarez & Gomez, 2016). Most of them emphasize describing the details of different citation function classification frameworks, datasets, and compression of experimental results.

Unlike these previous works, this survey focuses on the evaluation of the citation function classification task from both bibliometrics and modern NLP domains. We systematically review the research efforts in citation function classification in terms of annotation schemes, popular benchmarks, and state-of-the-art computational models over the past decades.

3. Methodologies

To conduct our literature overview, we adopted a systematic approach following the guidelines set by (Lyu et al., 2021; Iqbal et al., 2021). We began by compiling a list of candidate keywords, drawing from seed articles in our prior research on citation function classification. Using these keywords, we implemented a four-step strategy to ensure thorough retrieval of relevant studies. First, we searched for articles on citation function classification in the Web of Science (WOS) to identify appropriate search terms. This involved analyzing and recording words from titles, abstracts, and keywords. Based on this process, we formulated the following search terms: “citation function classification” OR “citing reason*” OR “citation motivation classification*” OR “citer motivation classification*” OR “citing motives classification” OR “citation purpose classification” OR “citation classif*” OR “citation taxonomy” OR “citation typology” OR “citation behavio*” OR “citation practice.” Additionally, we included broader concepts such as “citation behavior” and “citation practice” to capture a wider range of related studies.

Next, we conducted a literature search in September 2023 across three electronic databases—WOS, Google Scholar, and ACL Anthology—to encompass a broad spectrum of citation research. We also hand-searched reference lists of included articles and reviews to complement our search. Moreover, we tracked newly published studies throughout the research period to ensure no potential articles were omitted. This query yielded 108 publications indexed in the three databases. We then filtered out irrelevant publications by reviewing their abstracts and applied pre-processing and data screening techniques to refine the dataset, ultimately identifying the most pertinent publications. Finally, we conducted a comprehensive full-text review and comparative analysis of the selected research publications based on feature sets and accuracy. This process resulted in a core set of 31 key publications. Notably, we observed a slight increase in journal articles in recent years among these selected publications.

In this literature review, we particularly focus on new advancements leveraging deep learning approaches in citation function classification, topics that were not emphasized in previous reviews. Our comprehensive keyword-searching strategy ensures that we have included all necessary papers for the literature review. However, a limitation of our approach is the lack of an assessment of the risk of bias in the included studies.

4. Evolution of citation function classification

We follow the term “Citation Function”, introduced by Teufel et al. (2006b) to describe the different roles or purposes that citations play in scientific literature. This is equivalent to the terms used in other works, i.e., “Citing Motivation” (Lyu et al., 2021), “Citation Purpose” (Abu-Jbara et al., 2013), and “Citation Intent” (Roman et al., 2021). The fundamental idea behind them remains the same – citations in a scientific document can be categorized based on their textual content. We do not specifically discuss the sentiment or polarity of citations (Yousif et al., 2019b). Rather, we focus on identifying and classifying the different functions that a reference can serve.

Fig. 1 depicts the development process of citation function classification methods over time. It shows that different representative works have been proposed in different periods. Before the advent of computational model based automated classification, most scholars have to manually count the citations to quantify the different citation functions (Moravcsik & Murugesan, 1975; Chubin & Moitra, 1975; Oppenheim & Renn, 1978). By counting different citation functions, researchers could understand the *fine-grained* reasons why a particular paper is cited and the exact content it is cited for. However, manually labeling the citation function is

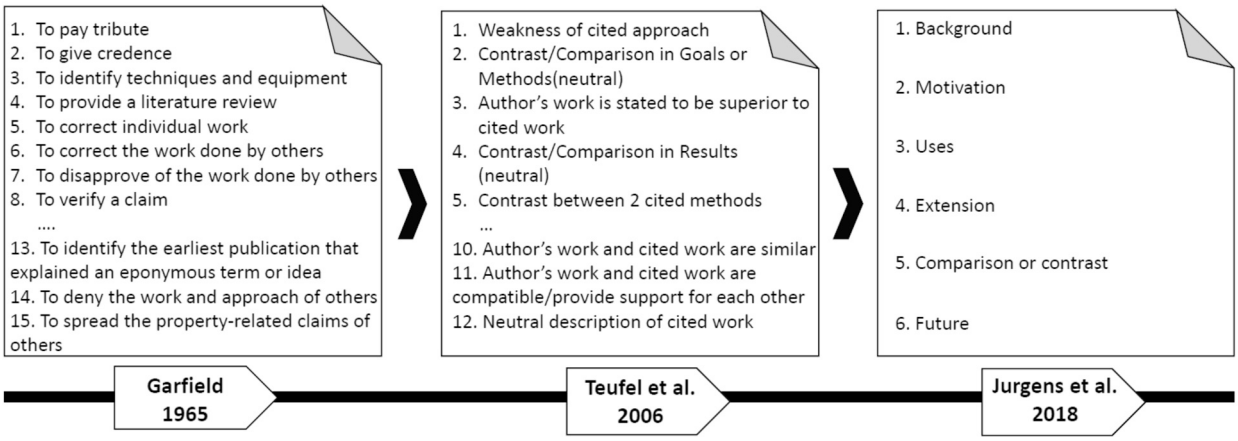

Fig. 2. Evolution of citation function classification annotation scheme from highly cited works of both bibliometrics and computer science domain.

time-consuming (Iqbal et al., 2021). With the increasing volume of scientific publications, the need for automated citation function classification has become increasingly important and practical. Machine learning, i.e., Support Vector Machine (SVM) (Joachims, 1998), has been employed in citation function classification to obtain impressive performance (Teufel et al., 2006b; Jurgens et al., 2018). In recent years, an array of intricate and sophisticated deep learning architectures has been progressively deployed in the execution of this task (Cohan et al., 2019). More recently, the use of pre-trained language models, such as BERT and SciBERT, has become increasingly popular in the citation function classification task and has demonstrated superior performance (Beltagy et al., 2019; Zhang et al., 2021; Roman et al., 2021). Based on the literature, we summarize the citation function classification task into three steps:

- Elaboration of the Citation Function Annotation Scheme;
- Construction of Citation Function Annotation Benchmarks; and
- Using training data from annotation benchmarks to train the automatic citation function classification models and evaluate the models' performance.

In most previous works, the citation function classification task is defined as a single-label classification task. We further formulate the citation function classification task in a machine-learning framework with mathematical notations. Let C denote the citation with a set of features, i.e., the text of the citation and the context in which it appears. Let Y denote a citation function label which is a label from a predefined set of classes, i.e., "Background information" or "Comparison". The mathematical formulation of the citation function classification task involves training a model to approximate a function $f : C \rightarrow Y$, where the model's objective is to accurately map each citation instance C to a corresponding label or function Y, based on the citation's contextual and content attributes. However, we note that some most recent research works also define citation function classification task as a multi-label classification task, since in the real scenario, one reference may serve for different functions of the citing paper (Lauscher et al., 2022; Zhang et al., 2021).

We further dive into the specific aspects of the citation function classification task by analyzing the change of annotation scheme in Section 4.1, different ways of constructing datasets in Section 4.2, and the evolution of computational models in Section 4.3. In each section, we describe the contributions of scholars in bibliometrics and computer science pertinent to the citation function classification task and how they learn from each other to further improve the task. In the meantime, by analyzing the remaining challenges of this task, we summarize how researchers from both disciplines can benefit from deeper collaboration in the future.

4.1. Annotation schemes

Garfield (1965) recognizes the significance of classification schemes for citations. Different citations may reflect different reasons and motivations for authors citing other people's works. MacRoberts and MacRoberts (1989) point out that citations have different types and should not be treated equally. By analyzing different types of citations, we can better understand an article.

Fig. 2 depicts how the citation function annotation scheme changes over time. We can see that the manual annotation frameworks proposed by early bibliometrics scholars are sociologically orientated, relatively complex, and difficult to distinguish (Teufel et al., 2006a). However, they are very detailed and academic (Garfield, 1965; Moravcsik & Murugesan, 1975). Nevertheless, the manual annotation scheme has some challenges as follows:

1. Unable to handle large volumes of literature (Li et al., 2013) – Traditional manual citation function annotation schemes require the annotator to understand all the selected articles and annotate all the citations' functions. As more and more articles are published online, it is impossible to label every article by hand.

2. Taking too much human effort (Pride & Knoth, 2017) – The annotator of the traditional manual citation function annotation scheme needs to be an expert in the specific research domain of the selected literature. The annotation process is time-consuming and monotony (Bertin et al., 2016).
3. The real scenarios of using citations can be complex (Lauscher et al., 2022; Zhang et al., 2022) – One cited paper can be used many times in one citing paper, and one citing sentence may contain multiple citations. These complex scenes will be a huge obstacle to manual annotation.

As a result, automatic citation function classification becomes crucial. Many computer scientists, inspired by bibliometricians, have designed automatic classification schemes. Teufel et al. (2006a) propose an automatic annotation scheme with twelve classes of citation functions inspired and modified from Spiegel-Rosing (1977). They remove one sociologically orientated class which is not easy to annotate in the real scenario. Jochim and Schütze (2012) fully inherit the four-faceted citation function classification proposed by bibliometric scholar (Moravcsik & Murugesan, 1975). The classification scheme can be summarized into four folds, i.e., conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, confirmative or negational, and each citation from the target corpus needs to be annotated with one function (Moravcsik & Murugesan, 1975). Dong and Schäfer (2011) also look into the "organic or perfunctory" dimension of Moravcsik and Murugesan (1975)'s scheme and further break it down into four classes: background, fundamental idea, technical basis, and comparison. Inspiring by Spiegel-Rosing (1977) and Teufel et al. (2006a), Jha et al. (2017) propose a six classes annotation scheme from a bibliometric perspective which later facilitated deep citation text analysis. Following Teufel et al. (2006a), Zhao et al. (2019) propose a new annotation scheme for defining the function of online resource citations.

We observe that, for some research domains such as bioinformatics, scholars also get inspiration from traditional manual annotation schemes of bibliometrics to design domain-specific automatic citation function classification schemes (Yu et al., 2009). Agarwal et al. (2010) develop an eight-class automatic annotation scheme for biomedical articles developed from Yu et al. (2009). After studying the manual annotation scheme of Jochim and Schütze (2012) and Teufel et al. (2006a), Abu-Jbara et al. (2013) conclude a six-class scheme: criticizing, comparison, use, substantiating, basis, and neutral (other). Also, citations are not equal, i.e., some of them can be influential (Zhu et al., 2015). Valenzuela et al. (2015) leverage the concept of citation function to recognize meaningful citations. A citation is considered meaningful if the citing paper uses or extends the ideas and information presented in the cited paper. Hassan et al. (2017) further extend the work of Valenzuela et al. (2015) to determine the importance of citations via citation functions. Jurgens et al. (2018) from the computer science community proposed a six-class annotation scheme and further leveraged the same scheme to measure the evolution of the NLP research domain via papers from ACL anthology (Jurgens et al., 2018). Authors always take the context and source of information into account when they cite the references. Lin (2018) from bibliometrics use an eight-class automatic annotation scheme to analyze the citation of articles representing six social science subjects. As a result, Lin (2018) finds that the distribution of citation functions across the six subject fields is significantly different, thereby indicating that there are variations in citing decisions among different disciplines in the Humanities and the Social Sciences.

Through the analysis of the previous literature, we have found a very interesting phenomenon, i.e., scholars from a bibliometrics background are inclined to design a complex and comprehensive annotation scheme which is more theoretical (Lyu et al., 2021). However, scholars from a computer science background tend to simplify the annotation scheme which is more practical (Hernández-Alvarez & Gomez, 2016; Pride & Knoth, 2017; Iqbal et al., 2021). Scholars from bibliometrics leverage the citation function classification to investigate the specific characteristics of their field which can facilitate a better understanding of the ideas, theories, and methods used in the discipline and how they relate to other disciplines. We also note that scholars from different research domains may modify the same previous citation function scheme in different ways. For example, both inspired by the six-class annotation of Jurgens et al. (2018), Cohan et al. (2019) simplify the scheme into three function classes. However, Zhang et al. (2021) develop a three dimension scheme with ten function classes. From a linguistic perspective, Bertin and Atanassova (2024) conducts an experiment on the semantic annotation of citation contexts in articles from seven PLOS journals, employing a rule-based approach and linguistic resources with ten distinct categories. Nevertheless, computer science researchers must account for not only the expense associated with manual annotation but also the capacity of the computational model to interpret the annotation scheme effectively, as this is critical for optimizing overall performance. In the future, scholars should find a balance between the two directions.

4.2. Datasets

Unlike other tasks, the goal of the citation function classification task requires high-quality annotated data. In most of the cases, the annotator needs to read and understand the citation sentences and their context to label the function. Previous researchers have conducted different ways to construct the dataset. Fig. 3 depicts the three ways of creating the citation function classification dataset:

(1) Expert annotation (Weinstock, 1971; Teufel et al., 2006b; Agarwal et al., 2010; Dong & Schäfer, 2011; Jurgens et al., 2018) – Following the tradition of bibliometrics, most previous works choose expert annotation since experts of a specific research domain have a deep understanding of the subject and are able to accurately identify the citation function based on their knowledge and experience. For example, Valenzuela et al. (2015) invite experts to annotate their benchmark. Normally, there will be an annotation guideline or a pilot annotation study to help the annotators to reach an agreement and make the decision before they formally start to annotate (Teufel et al., 2006a; Cohan et al., 2019). To ensure the quality of data, it is important to measure the inter-annotator agreement using Kappa coefficient (Cohen, 1960) or confidence score (Cohan et al., 2019). Expert annotation is the most common annotation method since it can help to create a consistent and standardized dataset of citation instances with citation function labels.

Table 2
Cross comparison of widely used citation function classification datasets.

Dataset	Domain(s)	Annotation	Total	Classes
Teufel (Teufel et al., 2006a)	Computer Linguistics	Expert	2,829	Neutral: 63%, Uses: 16%, Comparison in Goals: 4%, Similar: 4%, Weakness: 3%, Positive: 2%, Comparison in Results: 1%, Starting point: 1%, Contrast Method: 3%, Unfavourable Contrast: 1%, adapts: 1%, Support: 1%
DFKI (Dong & Schäfer, 2011)	Computer Science	Expert	1,768	Background: 65%, Technical Basis: 24%, Fundamental Idea: 7%, Comparison: 4%
UMICH (Jha et al., 2017)	Computer Linguistics	Expert	3,568	Criticizing: 16%, Comparison: 8%, Use: 18%, Substantiating: 8%, Basis: 5%, Neutral: 45%
Bakhti (Bakhti et al., 2018a)	Computer Science	Semi	2,000	Based on, Useful, Acknowledge, Contrast, Weakness, Hedges
ACL-ARC (Jurgens et al., 2018)	Computer Linguistics	Expert	1,989	Background: 51%, Extends: 4%, Uses: 19%, Motivation: 5%, Compare/Contrast: 18%, Future Work: 4%
SciCite (Cohan et al., 2019)	Computer Science & Medicine	Expert	11,020	Background: 58%, Method: 29%, Result Comparison: 13%
Act (Pride & Knoth, 2020)	Multiple Scientific Domains	Author	11,233	Background: 55%, Uses: 16%, Compare: 12%, Motivation: 10%, Extension: 6%, Future: 2%
TDMCite (Zhang et al., 2021)	Computer Linguistics	Expert	9,594	Motivation: 1%, Comparison: 8%, Extension: 1%, Application: 15%, Background: 28%, Method: 30%, Data: 5%, Result: 4%, Positive: 5%, Negative: 3%

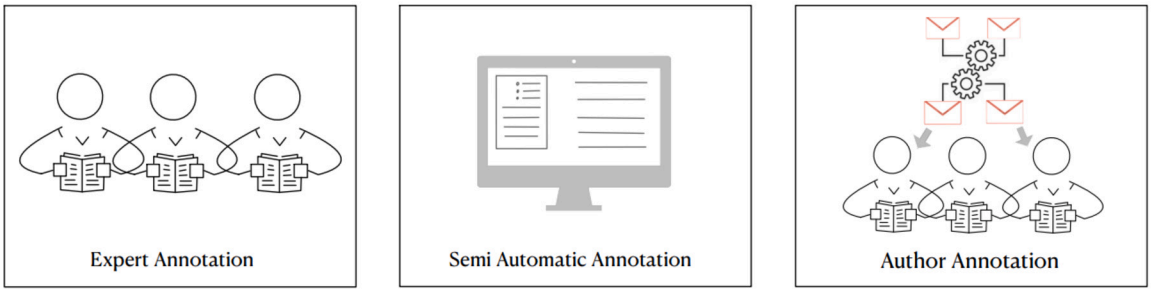

Fig. 3. Various methods for annotating data in the citation function classification task.

(2) Semi-automatic annotation (Bakhti et al., 2018a) – Previous studies have primarily relied on manual annotation methods to classify citations. This can be both time-consuming and dependent on specific subject domains. The use of semi-automatic annotation methods can help reduce time and human effort. However, this method still requires an accurate, manually labeled corpus and there is still room for improvement in automatic annotation.

(3) Author self-annotation (Bonzi & Snyder, 1991; Shadish et al., 1995; Pride & Knoth, 2020) – Questionnaire surveys, wherein the author of the citing paper is asked to annotate the citation's function, is the most classical method for collecting data on a citation's function (Vinkler, 1987; Case & Miller, 2011). The authors have a better understanding of their own works, and it is easier for them to identify the function of citations made by others (Case & Higgins, 2000). Recently, an Academic Citation Typing (ACT) platform has been implemented by scholars from the bibliometrics domain (Pride et al., 2019). This platform can send a survey containing a couple of questions to the original author of the citation and let the author of the citation annotate the function of their own publications into six classes for the citing paper. However, authors may have their own subjective ideas leading to inaccuracies in the annotation, and the response rate of the authors may be low.

The success of the citation function classification task is fundamentally contingent upon both the quality and the volume of the data utilized. Table 2 presents the details of popular datasets of citation function classification task. Firstly, as we can observe in the citation function classification task, whether a dataset is constructed by individual researchers, research institutions, or companies, the scale of the dataset is limited. For example, the famous benchmark, Acl-ACR citation dateset (Jurgens et al., 2018), has 1,989 labeled instances and the other commonly used benchmark, Scicite (Cohan et al., 2019), has 11,020 labeled instances. As we can observe from Table 2, the data sparsity problem lies in the citation function classification task. Most datasets have only a few thousand pieces of labeled data, while a few have more than tens of thousands of labeled data (Jurgens et al., 2018; Cohan et al., 2019). The scale of the manually annotated datasets is small because manual labeling requires sufficient time, expertise, and human effort.

As the statistics of the proportion of different citation function classes show, the data imbalance problem is quite severe. All of the presented datasets have imbalance issues, and this may be associated with the natural distribution of citations (Lyu et al., 2021; Zhang et al., 2021). For example, one paper may only have one citation that serves the inspiration purpose but many citations may contribute to the background knowledge. In fact, in most of the scientific literature, the related work section, compared with any other section, occupies the highest number of citations, and citations in this section have a high chance to suggest background knowledge (Lyu et al., 2021). For the above-discussed data scale issue and data imbalance problem, we also propose the solutions in Section 6. We observe that different datasets may have classes that overlap or are very similar to each other. For example, the class,

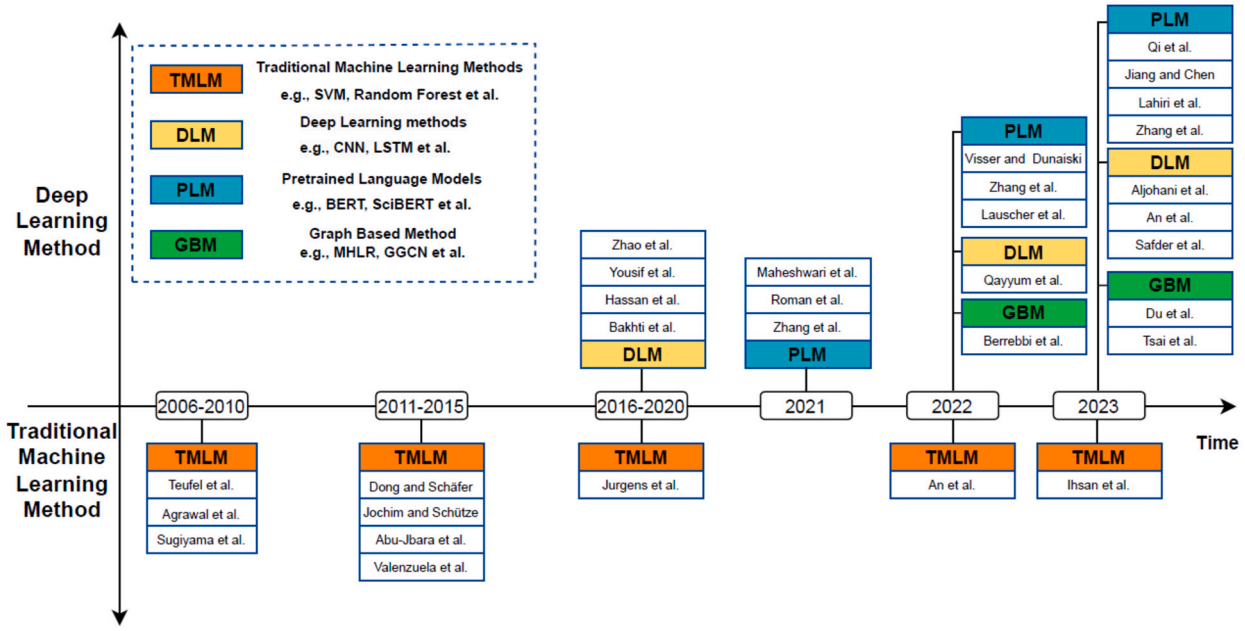

Fig. 4. This figure portrays the computational methods utilized in different research works on citation function classification across various time periods.

Idea, from Dong and Schäfer (2011) is similar to the class, Motivation, in the work of Jurgens et al. (2018). Most citation function classification datasets have the Background class, the Comparison class, and the Method class (Cohan et al., 2019).

Rules and guidelines for conducting dataset process annotations are crucial, as they determine the quality of the scheme and significantly impact the final performance of the classification task (Teufel et al., 2006a). In most previous studies, the rules and guidelines for citation function classification datasets are often descriptions of the citation function classes (Pride & Knoth, 2017). Having detailed and clear descriptions helps annotators understand the different citation function classes, thereby improving the efficiency and quality of the annotation process. We further observe two typical characteristics in these descriptions:

(1) Descriptions with Varying Levels of Detail: The descriptions can range from a few keywords to several sentences that explain the meaning of different citation function classes. For instance, Teufel et al. (2006b) described the "Use" class as "uses tools/algorithms/data/definitions." However, the study by Pride and Knoth (2020) described the Use class as "The citing paper uses the methodology or tools created by the cited paper."

(2) Descriptions with Diverse Descriptive Approaches: While different studies may use various methods to describe the same category, the underlying meaning generally remains consistent. For example, the study by Cohan et al. (2019) described the "Background" class as "The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field." However, the study by Jurgens et al. (2018) described the Background class as "P provides relevant information for this domain."

4.3. Classification models

As data mining and NLP technology continue to advance, more and more powerful computational models are being utilized in the citation function classification task. The recent availability of full-text scientific publications has enabled researchers to extract a vast amount of textural information from online resources. Like other text classification tasks, precision, recall, and F1-score are the most popular evaluation metrics in the citation function classification tasks. Fig. 4 delineates the chronological progression of computational methods employed in the domain of citation function classification over different time periods. It is observed that with the advancement and increasing sophistication of deep learning technology and pre-trained language models, there has been a notable shift towards the adoption of deep learning-based approaches in recent research endeavors pertaining to citation function classification tasks. In addition, Table 3 presents a more comprehensive comparison of various classification models and the outcomes of the citation function classification task.

Due to the interdisciplinary nature of this task, we have made every effort to compare different studies using the same datasets and metrics. To ensure thoroughness, we have also included additional papers that employ various evaluation metrics.

Traditional quantitative analysis – Previous research on citation function classification primarily relies on reading and understanding the literature by humans and further manually analyzing the citation content (Oppenheim & Renn, 1978; Frost, 1979). By sending the survey out and interviewing authors, bibliometrics scholars collect the data of citation function task and further manually calculate the count of different citation function classes (Vinkler, 1987; Case & Miller, 2011; Fazel & Shi, 2015).

Table 3
Computational classification models and results comparison. C.S. mean Computer Science and AAN means ACL Anthology Network dataset.

Studies	**Handcraft Feartures**	Nueral Network	Pretrained Lanaugae Model	Citation Network	Dataset	Data Repository	Number of Instances	Number of Classes	Best Performance
Teufel et al. (2006b)	Y	N	N	N	Teufel2006	Computational Linuistics	2,829	11	Marco-F1 0.57
Agarwal et al. (2010)	Y	N	N	N	GENIA Corpus	Biomedical	2,977	8	F1 0.75
Sugiyama et al. (2010)	Y	N	N	N	ACL-ARC (2007)	Computational Linguistics	10,921	2	Accuracy 0.88
Dong and Schäfer (2011)	Y	N	N	N	DFKI	Computational Linguistics	1,768	4	Marco-F1 0.66
Jochim and Schütze (2012)	Y	N	N	N	ACL-ARC (2004)	Computational Linguistics	8,032	8	Marco-F1 0.62
Abu-Jbara et al. (2013)	Y	N	N	N	AAN	Computational Linguistics	3,500	6	Marco-F1 0.58
Valenzuela et al. (2015)	Y	N	N	N	AAAI 2015	Computational Linguistics	450	2	Precision 0.62
Jurgens et al. (2018)	Y	N	N	N	ACL-ARC (Citation)	Computational Linguistics	1,969	6	Marco-F1 0.53
Bakhti et al. (2018a)	N	CNN	N	N	AAN	Computational Linguistics	8,700	6	Accuracy 0.65
Hassan et al. (2018a)	Y	LSTM	N	N	AAAI 2015	Computational Linguistics	450	2	Accuracy 0.92
Yousif et al. (2019a)	N	RCNN	N	N	AAN	Computational Linguistics	3,568	6	Marco-F1 0.85
Zhao et al. (2019)	N	LSTM	N	N	SciRes	C.S. & Medicine	3,087	6	Marco-F1 0.72
Roman et al. (2021)	N	N	BERT	N	SciCite	C.S. & Medicine	11,020	3	Precision 0.89
Maheshwari et al. (2021)	N	Bi-LSTM	SciBERT	N	3C Shared Task 2021	Computer Science	3,000	6	Macro-F1 0.68
Zhang et al. (2021)	N	N	BERT	N	TDMCite	Computer Science	9,594	10	Marco-F1 0.60
An et al. (2022)	Y	N	N	N	Semi (dataset 2)	10 Different Disciplines	2,685	2	AUC-ROC 0.96
Qayyum et al. (2022)	Y	MLP-ANN	N	N	AAAI2015	Computational Linguistics	450	2	Precision 0.82
Visser and Dunaiski (2022)	N	N	SciBERT	N	SciCite	C.S. & Medicine	11,020	3	F1 0.85
Lauscher et al. (2022)	N	N	SciBERT & RoBERTa	N	MultiCite	Computational Linguistics	12,653	7	Strict Accuracy Scores 0.70
Berrebbi et al. (2022)	N	MLP	SciBERT	GAT	SciCite	C.S. & Medicine	11,020	3	Macro-F1 0.89
Zhang et al. (2022)	Y	LSTM	BERT T5	N	Ni-Cite	C.S. & Medicine	11,195	3	Marco-F1 0.85
Ihsan et al. (2023)	Y	N	N	N	AAN	Computational Linguistics	7,390	8	F1 0.91
Aljohani et al. (2023)	N	CNN	N	N	ACL Articles	Computational Linguistics	9,518	6	Accuracy 0.72
An et al. (2023)	Y	CNN	N	N	AAN	Computational Linguistics	456	4	AUC-ROC 0.55
Safder et al. (2023)	N	Bi-LSTM	N	N	ACL Articles	Computational Linguistics	9,518	6	F1 0.80
Qi et al. (2023)	Y	Bi-LSTM	SciBERT	N	SciCite	C.S. & Medicine	11,020	3	Macro-F1 0.83
Jiang and Chen (2023)	Y	MLP	SciBERT	N	Jiang2021	Computational Linguistics	4,784	6	Macro-F1 0.74
Lahiri et al. (2023)	N	N	GPT-2 SciBERT	N	SciCite	C.S. & Medicine	11,020	3	Macro-F1 0.86
Du et al. (2023)	N	MLP	N	MHLR	SciCite (resplit)	C.S. & Medicine	5,766	3	Macro-F1 0.62
Tsai et al. (2023)	Y	N	Longformer	GGCN	MultiCite	Computational Linguistics	12,653	7	Strict Accuracy Scores 0.71
Zhang et al. (2023a)	N	N	GPT-2	N	TDMCite	Computer Science	9,594	8	Marco-F1 0.67

Machine learning with handcrafted features – As increasing numbers of scientific publications emerge online, the automatic recognition of the function of citations becomes essential (Zhang et al., 2022). Teufel et al. (2006b) is the pioneering work that utilizes the traditional machine learning classifier (IBK classifier) with different features, including the cue phrases. Several subsequent studies have utilized different machine learning classifiers with different features in citation function classification tasks. Abu-Jbara et al. (2013) utilize a few machine learning classifiers in citation function classification tasks and SVM obtains the best performance with the proposed structural and lexical features. Following Teufel et al. (2006b) and Abu-Jbara et al. (2013), Jha et al. (2017) employ three classical machine learning classifiers, i.e., SVM, Naive Bayes, Random Forest in citation function classification with the same

set of features proposed by Abu-Jbara et al. (2013). Valenzuela et al. (2015) suggest an SVM-based classification method designed to tackle this task by employing a comprehensive set of features. These features span a variety of aspects, from the number of citations to the specific locations where these citations are found within the text of the paper. Abu-Jbara et al. (2013) and Jha et al. (2017) show that lexical and structural features are essential in citation function classification. Siddharthan and Teufel (2007) conduct five different basic machine learning classifiers in their work and prove that scientific attribution features and lexical, linguistic, and position-based features are all helpful in improving the classification performance of the models. Dong and Schäfer (2011) utilize the Naive Bayes and Bayes Network as their classification model, and present textual, physical, and syntactic features (part-of-speech) as new features. Jochim and Schütze (2012) and Jurgens et al. (2018) create meaningful features that help to improve the classification performance of the citation function classification task. Linguistic features such as n-grams (Agarwal et al., 2010; Sugiyama et al., 2010) and the cue words (Wang et al., 2012) have also been widely used. From their experimental results, we observe that features like uni-gram and n-gram are more useful when dealing with large-scale datasets. We also note that SVM and Naive Bayes are the most popular and robust classical machine learning models in citation function classification before the emergence of the neural network (Dong & Schäfer, 2011; Abu-Jbara et al., 2013; Jha et al., 2017; Jurgens et al., 2018). Furthermore, traditional machine learning models can be employed in a self-training approach, harnessing the unlabeled instances of the dataset to enhance the performance of citation function classification models, as noted by (An et al., 2022). Recently, Ihsan et al. (2023) employed a suite of machine-learning models on an annotated dataset, including the Support Vector Machine (SVM), Naïve Bayesian (NB), and Random Forest (RF). They incorporated various part-of-speech elements (e.g., Nouns, Verbs, Adverbs, and Adjectives) as innovative features. This approach achieved an impressive accuracy of 91%.

Deep learning models – With the advent of deep learning, the approaches to citation function classification task have taken a shift. Bakhti et al. (2018b) incorporate the text representation of citation context into a convolutional neural network (CNN) (Gu et al., 2018) model to classify citations into different citation functions. Following a similar strategy, Aljohani et al. (2023) utilizes a CNN model for citation function classification. Diverging from traditional CNN classification models, they integrate their CNN with fastText-based pre-trained embedding vectors. Remarkably, using only the citation context as input, they surpass performance in both binary (distinguishing between important and non-important) and multi-class citation classification tasks. In addition, Yousif et al. (2019a) employ Recurrent Convolutional Neural Networks (RCNN) with a multi-task learning framework. Following a similar path, An et al. (2023) use a generative model to highlight key citations. They extract features from a topic model and the Citation Influence Model (CIM), and combine them with 13 other standard features. These are then processed through SVM, Random Forest (RF), and a CNN to identify the most significant citations. Hassan et al. (2018a, 2018b) employ the LSTM model (Hochreiter & Schmidhuber, 1997) into citation function classification task. They further prove that deep learning models can obtain better performance than the traditional machine learning in citation function classification tasks with large scale dataset. Zhao et al. (2019) also leverage the LSTM model to build the multi-task learning framework with classification task and recommendation task. Continuing in this vein, Cohan et al. (2019) introduced Structural Scaffolds, a novel multi-task learning framework that leverages the BiLSTM model alongside attention mechanisms.

In contrast to the complex neural network models designed for multi-class citation function classification, the Multilayer Perceptron Artificial Neural Network (MLP-ANN) has also been employed as a binary classifier specifically for identifying important citations, as highlighted by Qayyum et al. (2022). Recently, Safder et al. (2023) propose a deep learning architecture for citation context classification. Intriguingly, they transform the formal machine translation task into a citation classification challenge. Distinct from the feature-based state-of-the-art models, their innovative focal-loss and class-weight-aware BiLSTM approach, combined with pre-trained GloVe embedding vectors, solely relies on the citation context as input. This novel strategy leads them to outperform existing models in multi-class citation context classification tasks.

We note that scholars from computer science are trying their best to introduce novel deep learning models into citation function classification tasks. We note that after neural networks became popular, researchers choose to use neural networks to build their computational models and employ traditional machine learning models, i.e., SVM, as one of their baseline models.

Pre-trained language models – In contemporary NLP, the field has witnessed significant advancements through the utilization of pre-trained language models, such as ELMO (Peters et al., 2018), the GPT series (Radford et al., 2018, 2019; Brown et al., 2020), and BERT series (Devlin et al., 2019; Beltagy et al., 2019). These models are founded upon the principles of pre-training on extensive corpora, resulting in notable improvements in their ability to perform a wide range of NLP tasks. This progress has not gone unnoticed, as researchers have rapidly adopted these techniques in various domains, including citation function classification (Visser & Dunaiski, 2022).

For instance, Roman et al. (2021) leverage the power of BERT to incorporate textual representations of citation contexts within the citation intent classification task, achieving superior performance when compared to baseline models. This success has set a precedent for further exploration in the same direction. Following this trend, Zhang et al. (2021), Lauscher et al. (2022), and Zhang et al. (2022) have also chosen to employ BERT for citation function classification due to its demonstrated capacity to capture nuanced textual representations, ultimately enhancing classification performance.

We note that researchers also integrate pre-trained language with other deep learning models to construct a powerful computational framework. Cohan et al. (2019) integrate ELMo and attention-based LSTM model to obtain better text representation. Additionally, there exists the fusion of a deep learning model with a persistent language model. Qi et al. (2023) employ a multi-task learning framework that builds upon a Bi-LSTM architecture and utilizes the SciBERT encoder to achieve enhanced classification performance. Lauscher et al. (2022) and Maheshwari et al. (2021) employ the SciBERT (Beltagy et al., 2019) model into their classi-

fication computational models. Jiang and Chen (2023) introduce a task focused on the contextualized segment-wise classification of citation functions. They also present an array of robust classification models based on SciBERT for this purpose. In the most recent work of Zhang et al. (2022), a novel end-to-end model T5 has also been utilized in citation function classification (Raffel et al., 2020). Recently, state-of-the-art large language models have been brought into play within this research domain. Zhang et al. (2023a) utilized the GPT-2 Radford et al. (2019) large language model to develop HybridDA, a two-stage model that integrates data augmentation and data retrieval. The model aims to generate higher quality annotated data for citation function, addressing the challenges of data imbalance and data sparsity issues for citation function classification task. A novel approach to utilizing generative large language models in citation function classification tasks involves prompt learning. Shui et al. (2024) combine pretrained language models with a multi-task learning (MTL) framework for citation intention classification, achieving strong performance. Lahiri et al. (2023) unveiled CitePrompt, a GPT-2-based framework that adopts this innovative, yet unexplored method for classifying citation functions.

Graph-based approaches – Beyond the previously discussed methods, graph-based citation function classification techniques have also surfaced recently. Du et al. (2023) introduce an innovative text-free approach for citation intent classification. Leveraging a knowledge graph anchored on the SciCite dataset, their method proficiently extracts citation details from publications, paving the way for precise citation intent predictions. Central to their strategy is the collection of weakly labeled data, which culminates in the formation of a robust, large-scale knowledge graph. Furthermore, their analysis spans both transductive and inductive frameworks, ensuring a well-rounded examination of the task. Berrebbi et al. (2022) introduce GraphCite, offering a fresh angle on the task of citation function classification. Rather than solely relying on textual hints within the citation phrase, their approach integrates the citation graph, tapping into the intricate patterns of citation connections. Within this innovative framework, they meticulously assessed the effectiveness of graph-based models in predicting function. In a related vein, Tsai et al. (2023) have developed a comprehensive citation graph, which labels citation intents and their associated supporting evidence between citing and cited papers. Their model utilizes SciBERT for processing a wide array of papers. In order to understand the coreference relationships among words and sentences within a paper, they have implemented the Gated Graph Convolution Network (GGCN), which is used to construct a coreference graph. We observe that all the aforementioned studies have incorporated pre-trained language models into their Grahp-based deep-learning model architectures (Chen et al., 2023).

In summary, given a citation function classification task, one should follow these principles to choose the appropriate approaches: (1) For datasets containing diverse textual data relevant to the task or when managing multiple tasks, employing deep learning-based methods is recommended; (2) In cases where simulating real-world citation scenarios is essential, the use of a range of pre-trained language models is advised; (3) When confronted with small or imbalanced datasets, the application of generative models is suggested as an effective strategy; and (4) If the dataset includes citation network information, adopting a graph-based approach is a prudent choice.

5. Applications of citation function classification

Citation function classification can facilitate many citation content-based applications and research. Firstly, it helps to create intelligent databases of scientific publications such as Semantic Scholar.[1] Semantic Scholar employs the citation function in scientific publications to assist researchers in more effectively locating and comprehending references. Secondly, it enables better evaluation and comparison of different research domains. For example, creating a fine-grained citation indexer based on different citation functions becomes possible. Therefore, in this section, we will explore the implications of our proposed approaches by highlighting their various future application scenarios.

5.1. Citation summarization

In a variety of text-mining fields, citations are acknowledged as a significant information source for the generation of automatic summaries (Yasunaga et al., 2019; Zaib et al., 2022, 2023). Citation-based summarization utilizes text from multiple researchers to identify the key aspects of a target paper. Prior research in this area has primarily focused on the extraction aspect of the task, which involves selecting a set of citation sentences that effectively showcase the contribution of the target paper (Ding et al., 2014). More recently, some researchers have leveraged the state-of-the-art pre-trained language models for the automatic generation of citation texts. However, they are still focused on sentence-level citation text generation (Xing et al., 2020). In actual writing situations, it is common for authors to summarize several studies in a single sentence or discuss pertinent information across an entire paragraph. Furthermore, previous research has identified multiple citation functions indicating that writers may require control over the intended meaning of generated sentences to suit various scenarios (Wu et al., 2021). Generating a citation sentence based on the citation function can be more suitable for different scenarios. Our models can function as a part of a multi-task framework for citation summarization. The optimal approach to using our models in the citation summarization task would involve following the same fine-tuning strategy, thereby making the pre-trained language model familiar with the citation function's meaning and generating various citation sentences with different citation intents.

[1] https://www.semanticscholar.org/.

5.2. Citation recommendation

The task of citation recommendation involves suggesting relevant citations for a particular text (Jebari et al., 2023). With the substantial growth of scientific literature, conducting a comprehensive literature review has become increasingly difficult despite significant progress in digital libraries and information retrieval systems. Citation recommendation is a useful tool in enhancing the quality and efficiency of this process by suggesting relevant scientific publications as potential citations for a given query document. Nevertheless, achieving high-quality citation recommendations is a challenging task. It requires not only suggesting relevant citations for the paper being composed but also ensuring that the recommended citations are contextually appropriate for the specific locations where they are referenced. To address these challenges, different deep learning models have been developed that recommend research papers to users with a focus on providing personalized recommendations. Other researchers propose a novel citation recommendation task which aims to recommend a diverse citation context list based on all the citation context sentences extracted from a list of citing articles (Chen et al., 2020). Generally, citation recommendation systems can be categorized into three primary approaches: content-based methods, collaborative filtering, and graph-based methods. For example, the models can serve as a classifier to find citations with particular functions.

5.3. Evaluating the evolution of a scientific field

The expansion of the scientific research communities has resulted in the emergence of new publication venues. It is worth investigating whether these new venues have become institutionalized, i.e., whether they resemble established conferences or have developed their own unique style and captured different representations of knowledge. The evolution of new publication venues can be observed through citation framing and which measures the extent to which the framing of papers in a newer venue resembles that of papers in established venues (Jurgens et al., 2018). As scientific fields evolve, new subfields often emerge around specific methods or technologies and become the focus of collective problem-solving and continuous improvement. This can be observed in the increasing number of citation functions related to “USE” and “Methodology” over time. The process of evaluating the evolution of a scientific field involves three key steps. Firstly, it is necessary to design a citation annotation framework. Secondly, the citation function of data from the target research domain needs to be classified. Finally, the classification output should be quantitatively analyzed. In light of this, the multi-aspect citation function annotation framework can be applied to evaluate the development of a scientific field as it is designed with a document-level perspective.

6. Findings and challenges

6.1. Findings

Firstly, we find that in the case of citation function classification tasks, bibliometrics and the computer science research community benefit each other. Scholars from computer science often utilize traditional citation function annotation schemes from bibliometrics research in their work (Teufel et al., 2006b). Scholars from a bibliometrics background tend to use computational models to realize automatic citation function classification (Small, 2018). In general, scholars from computer science concentrate more on improving the computational model to obtain better performance. However, scholars from a bibliometrics background are more interested in utilizing the automatic classification method for citation analysis downstream tasks (Jurgens et al., 2018). Secondly, scholars who are interested in citation function classification always follow and utilize the cutting-edge techniques from NLP and data mining (Cohan et al., 2019; Lauscher et al., 2022; Zhang et al., 2022). In addition, the collaboration between bibliometrics and computer science can be further deepened.

6.2. Major problems

As delineated in Section 4.2, the citation function classification task is characterized by issues of data imbalance and data sparsity. The data imbalance issue is an inherent challenge, attributable to the uneven natural distribution of various citation types within a single scientific literature, as highlighted by Lyu et al. (2021). To address the issue of class imbalance, Jha et al. (2017) proposes the exclusion of the neutral class, noting that it encompasses more than half of the citations. However, the latest text generation model, i.e., GPT-3, has been proven to be able to generate text that is similar to the text written by humans. It is possible for researchers to utilize the text generation model to create more annotated data to fill in the imbalanced classes or tail classes. However, how to better leverage synthetic data in citation function classification remains uncertain. The data sparsity problem is caused by the high cost of manual labelling in citation function classification. For example, to investigate a new research domain, it is not easy to obtain high-quality annotated data on a large scale. For low resource setting tasks, Few-shot Learning provides a new solution (Zhao et al., 2022). It is possible for researchers to utilize the Few-shot Learning model (Wang et al., 2020) to deal with small-scale datasets on citation function classification tasks and to obtain a relatively good performance. Also, most of the datasets in this task are domain-specific, and it is not easy to directly compare the models between different works.

7. Possible future directions

The new NLP techniques in terms of computational approaches can also be explored to enhance the performance of citation function classification tasks. Several cutting-edge NLP techniques are yet to be explored in the task of citation function classification. This, therefore, leads to several directions for future research expansions.

7.1. Large language model

In recent times, pre-trained language models have gained significant attention as they have shown remarkable proficiency in solving various NLP tasks. Pre-trained language models, i.e., BERT with small-scale parameters, may not be as powerful as the large language models with billions of parameters like GPT-3 and GPT-4. This phenomenon can be attributed to the observation that when language models scale beyond a certain threshold in terms of parameters, they not only demonstrate considerable enhancements in performance but also manifest unique capabilities absent in smaller-scale language models. An impressive implementation of the Large Language Model is the ChatGPT which utilizes Large Language Models from the GPT series for dialogue and showcases remarkable conversational abilities with humans. In consideration of the aforementioned, the employment of a powerful generative language model can be advantageous for the citation function classification task in two distinct ways. Firstly, the traditional deep learning approach can be improved by replacing GPT-2 with a more robust generative language model such as GPT-4. This may result in the synthesis of higher-quality training data for the citation function classification task. Secondly, it is feasible to transform the citation function classification task into a sequence-to-sequence task by allowing the generative language model to generate citation functions based on citation context input. This may ultimately enhance the classification performance (Zhang et al., 2023b).

7.2. Prompt learning

A paradigm shift is underway concerning the adaptation of prompt learning of large language models for citation function classification (Lahiri et al., 2023). This shift originated from the studies conducted on T5 and GPT-3 and revealed that pre-trained language models could be efficiently activated through the utilization of textual prompts or demonstrations. The advent of prompt learning has emerged as a novel paradigm within contemporary NLP. This approach involves directly adapting pre-trained language models to tasks such as cloze-style prediction, autoregressive modeling, or sequence-to-sequence generation, marking a significant development in the field. This ultimately leads to promising performances across various tasks. The fundamental concept behind prompt learning is to integrate a template into the input which effectively transforms text classification tasks into equivalent cloze-style tasks (Zhu et al., 2022). A further direction of using prompt learning in citation function classification task may result in prompt trainer engineering, i.e., designing a better prompt-based tuning strategy for the citation function classification task.

7.3. Instruction tuning

In the realm of NLP, instruction tuning has become a groundbreaking approach, leveraging natural language instructions in tandem with language models to enable zero-shot performance on tasks that have not been previously encountered. The idea of instruction tuning is introduced by the team led by Quoc V. Le at Google Deepmind. This concept involves generating individual instructions (hard tokens) for each task and unfreezing the pre-trained model parameters followed by fine-tuning on multiple full-shot tasks. Finally, the model's generalization ability (zero-shot) is evaluated for the specific task. Rosenbaum et al. (2022) showcase the efficiency of instruction fine-tuning through their innovative method, LINGUIST, employing the large-scale seq2seq model, AlexaTM 5B, for directing the outputs in multilingual intent- and slot-labeled data generation. They were the pioneers in this demonstration. Keeping this in mind, we can enhance the generative model by utilizing instruction tuning to address the low-resource challenges in citation function classification tasks.

8. Conclusion

This paper takes the first step toward summarizing existing research efforts on citation function classification from an interdisciplinary perspective. We conduct a systematic and comprehensive study on different aspects of citation function classification tasks pertinent to annotation schemes, benchmarks, and computational models. Finally, we summarize the limitations of existing research works and propose several promising research directions. Our work has the potential to benefit researchers from both computer science and bibliometrics backgrounds. The limitation of our work is that we only include the primary studies published in the English language. However, numerous notable bibliometric studies on citation functions have been published in languages other than English, e.g., Japanese.

CRediT authorship contribution statement

Yang Zhang: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yufei Wang:** Writing – review & editing, Writing – original draft, Methodology. **Quan Z. Sheng:** Writing – review & editing, Funding acquisition. **Lina Yao:** Writing – original draft, Funding acquisition. **Haihua Chen:** Writing – review & editing, Methodology. **Kai Wang:** Writing – review & editing, Writing – original draft. **Adnan Mahmood:**

Writing – review & editing. **Wei Emma Zhang:** Writing – review & editing. **Munazza Zaib:** Writing – review & editing. **Subhash Sagar:** Writing – review & editing. **Rongying Zhao:** Writing – review & editing, Conceptualization.

References

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 596–606).

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. In *Proceedings of AMIA annual symposium, American medical informatics association* (p. 11).

Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2023). A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations. *Journal of Information Science*, *49*, 79–92.

An, X., Sun, X., & Xu, S. (2022). Important citations identification with semi-supervised classification model. *Scientometrics*, *127*, 6533–6555.

An, X., Sun, X., Xu, S., Hao, L., & Li, J. (2023). Important citations identification by exploiting generative model into discriminative model. *Journal of Information Science*, *49*, 107–121.

Bakhti, K., Niu, Z., & Nyamawe, A. S. (2018a). Semi-automatic annotation for citation function classification. In *2018 international conference on control, artificial intelligence, robotics & optimization* (pp. 43–47).

Bakhti, K., Niu, Z., Yousif, A., & Nyamawe, A. S. (2018b). Citation function classification based on ontologies and convolutional neural networks. In *International workshop on learning technology for education in cloud* (pp. 105–115).

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3615–3620).

Berrebbi, D., Huynh, N., & Balalau, O. (2022). Graphcite: Citation intent classification in scientific publications via graph embeddings. In *Companion proceedings of the web conference 2022* (pp. 779–783).

Bertin, M., & Atanassova, I. (2024). Linguistic perspectives in deciphering citation function classification. *Scientometrics*, 1–13.

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *The Journal of the Association for Information Science and Technology*, *67*, 164–177.

Bonzi, S., & Snyder, H. (1991). Motivations for citation: A comparison of self citation and citation to others. *Scientometrics*, *21*, 245–254.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, *51*, 635–645.

Case, D. O., & Miller, J. B. (2011). Do bibliometricians cite differently from other scholars? *Journal of the American Society for Information Science and Technology*, *62*, 421–432.

Chen, C., Wang, Y., Zhang, Y., Sheng, Q. Z., & Lam, K. Y. (2023). Separate-and-aggregate: A transformer-based patch refinement model for knowledge graph completion. In *International conference on advanced data mining and applications* (pp. 62–77). Springer.

Chen, H., Yang, Y., Lu, W., & Chen, J. (2020). Exploring multiple diversification strategies for academic citation contexts recommendation. *Electronic Library*, *38*, 821–842.

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

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 3586–3596).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186).

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *The Journal of the Association for Information Science and Technology*, *65*, 1820–1833.

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of 5th international joint conference on natural language processing* (pp. 623–631).

Du, X., Ahrabian, K., Ananthan, A. B. S., Myloth, R. D., & Pujara, J. (2023). Graph-based structure aware citation intent classification. In *Proceedings of the AAAI 2023 workshop*.

Fazel, I., & Shi, L. (2015). Citation behaviors of graduate students in grant proposal writing. *Journal of English for Academic Purposes*, *20*, 203–214.

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

Garfield, E. (1965). Can citation indexing be automated? In *Statistical association methods for mechanized documentation: Symposium proceedings* (pp. 189–192).

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Hassan, S. U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *Proceedings of the 2017 ACM/IEEE joint conference on digital libraries* (pp. 1–8).

Hassan, S. U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., Ventura, S., & Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems*, *192*, Article 105383.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018a). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics*, *117*, 1645–1662.

Hassan, S. U., Iqbal, S., Imran, M., Aljohani, N. R., & Nawaz, R. (2018b). Mining the context of citations in scientific publications. In *Proceedings of international conference on Asian digital libraries* (pp. 316–322).

Hernández-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, *22*, 327–349.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hu, Z., Cui, J., & Lin, A. (2023). Identifying potentially excellent publications using a citation-based machine learning approach. *Information Processing & Management*, *60*, Article 103323.

Ihsan, I., Rahman, H., Shaikh, A., Sulaiman, A., Rajab, K., & Rajab, A. (2023). Improving in-text citation reason extraction and classification using supervised machine learning techniques. *Computer Speech & Language*, *82*, Article 101526.

Iqbal, S., Hassan, S. U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. *Scientometrics*, *126*, 6551–6599.

Jebari, C., Herrera-Viedma, E., & Cobo, M. J. (2023). Context-aware citation recommendation of scientific papers: Comparative study, gaps and trends. *Scientometrics*, 1–26.

Jha, R., Jbara, A. A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*, 93–130.

Jiang, X., & Chen, J. (2023). Contextualised segment-wise citation function classification. *Scientometrics*, *128*, 5117–5158.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of machine learning: ECML-98: 10th European conference on machine learning.* (pp. 137–142).
Jochim, C., & Schütze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of 24th international conference on computational linguistics* (pp. 1343–1358).
Jung, S., & Segev, A. (2013). Analyzing future communities in growing citation networks. In *Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing* (pp. 15–22).
Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391–406.
Lahiri, A., Sanyal, D. K., & Mukherjee, I. (2023). Citeprompt: Using prompts to identify citation intent in scientific papers. Preprint, arXiv:2304.12730.
Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Cohan, A., Jurgens, D., & Lo, K. (2022). Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1875–1889).
Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards fine-grained citation function classification. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013* (pp. 402–407).
Lin, C. S. (2018). An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. *Scientometrics*, *116*, 797–813.
Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: A meta-synthesis. *Scientometrics*, *126*, 3243–3264.
MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*, 342–349.
Maheshwari, H., Singh, B., & Varma, V. (2021). Scibert sentence representation for citation context classification. In *Proceedings of the second workshop on scholarly document processing* (pp. 130–133).
Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*, 86–92.
Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, *29*, 225–231.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 2227–2237).
Pride, D., & Knoth, P. (2017). Incidental or influential? - Challenges in automatically detecting citation importance using publication full texts. In *Proceedings of research and advanced technology for digital libraries: 21st international conference on theory and practice of digital libraries* (pp. 572–578).
Pride, D., & Knoth, P. (2020). An authoritative approach to citation classification. In *Proceedings of the 2020 ACM/IEEE joint conference on digital libraries* (pp. 337–340).
Pride, D., Knoth, P., & Harag, J. (2019). Act: An annotation platform for citation typing at scale. In *Proceedings of the 2019 ACM/IEEE joint conference on digital libraries* (pp. 329–330).
Qayyum, F., Jamil, H., Iqbal, N., Kim, D., & Afzal, M. T. (2022). Toward potential hybrid features evaluation using mlp-ann binary classification model to tackle meaningful citations. *Scientometrics*, *127*, 6471–6499.
Qi, R., Wei, J., Shao, Z., Li, Z., Chen, H., Sun, Y., & Li, S. (2023). Multi-task learning model for citation intent classification in scientific publications. *Scientometrics*, 1–21.
Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics*, *14*, Article 101047.
Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. https://blog.openai.com/language-unsupervised.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 9.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 5485–5551.
Roman, M., Shahid, A., Khan, S., Koubaa, A., & Yu, L. (2021). Citation intent classification using word embedding. *IEEE Access*, *9*, 9982–9995.
Rosenbaum, A., Soltan, S., Hamza, W., Versley, Y., & Boese, M. (2022). Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th international conference on computational linguistics* (pp. 218–241).
Safder, I., Ali, M., Aljohani, N. R., Nawaz, R., & Hassan, S. U. (2023). Neural machine translation for in-text citation classification. *The Journal of the Association for Information Science and Technology*, *74*, 1229–1240.
Safer, M. A., & Tang, R. (2009). The psychology of referencing in psychology journal articles. *Perspectives on Psychological Science*, *4*, 51–53.
Shadish, W. R., Tolliver, D., Gray, M., & Sen Gupta, S. K. (1995). Author judgements about works they cite: Three studies from psychology journals. *Social Studies of Science*, *25*, 477–498.
Shui, Z., Karypis, P., Karls, D. S., Wen, M., Manchanda, S., Tadmor, E. B., & Karypis, G. (2024). Fine-tuning language models on multiple datasets for citation intention classification. Preprint, arXiv:2410.13332.
Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of human language technologies 2007: The conference of the North American chapter of the association for computational linguistics* (pp. 316–323).
Small, H. (1982). Citation context analysis. *Progress in Communication Sciences*, *3*, 287–310.
Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, *12*, 461–480.
Spiegel-Rosing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, *7*, 97–113.
Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010). Identifying citing sentences in research papers using supervised learning. In *Proceedings of the 2010 international conference on information retrieval & knowledge management* (pp. 67–72).
Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial workshop on discourse and dialogue* (pp. 80–87).
Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 103–110).
Tsai, H. J., Yen, A. Z., Huang, H. H., & Chen, H. H. (2023). Citation intent classification and its supporting evidence extraction for citation graph construction. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 2472–2481).
Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *Proceedings of the workshops at the twenty-ninth AAAI conference on artificial intelligence* (p. 13).
Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics*, *12*, 47–72.
Visser, R., & Dunaiski, M. (2022). Sentiment and intent classification of in-text citations using bert. In *Proceedings of the 43rd conference of the South African institute* (pp. 129–145).
Wang, W., Villavicencio, P., & Watanabe, T. (2012). Analysis of reference relationships among research papers, based on citation context. *International Journal on Artificial Intelligence Tools*, *21*, Article 1240004.
Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*, 1–34.
Weinstock, N. (1971). Citation indexes. *Encyclopedia of Library and Information Science*, *5*, 16–41.
Wu, J. Y., Shieh, A. T. W., Hsu, S. J., & Chen, Y. N. (2021). Towards generating citation sentences for multiple references with intent control. Preprint, arXiv:2112.01332.

Xing, X., Fan, X., & Wan, X. (2020). Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6181–6190).

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R. (2019). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth AAAI symposium on educational advances in artificial intelligence* (pp. 7386–7393).

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019a). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, *335*, 195–205.

Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019b). A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*, *52*, 1805–1838.

Yu, H., Agarwal, S., & Frid, N. (2009). Investigating and annotating the role of citation in biomedical full-text articles. In *Proceedings of the 2009 IEEE international conference on bioinformatics and biomedicine workshop* (pp. 308–313).

Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems*, *64*, 3151–3195.

Zaib, M., Zhang, W. E., Sheng, Q. Z., Sagar, S., Mahmood, A., & Zhang, Y. (2023). Learning to select the relevant history turns in conversational question answering. In *International conference on web information systems engineering* (pp. 334–348). Springer.

Zhang, Y., Wang, Y., Sheng, Q. Z., Mahmood, A., Emma Zhang, W., & Zhao, R. (2021). TDM-CFC: Towards document-level multi-label citation function classification. In *Proceedings of the 22nd international conference on web information systems engineering* (pp. 363–376).

Zhang, Y., Wang, Y., Sheng, Q. Z., Mahmood, A., Zhang, W. E., & Zhao, R. (2023a). Hybrid data augmentation for citation function classification. In *2023 international joint conference on neural networks (IJCNN), IEEE* (pp. 1–8).

Zhang, Y., Wang, Y., Wang, K., Sheng, Q. Z., Yao, L., Mahmood, A., Zhang, W. E., & Zhao, R. (2023b). When large language models meet citation: A survey. Preprint, arXiv:2309.09727.

Zhang, Y., Zhao, R., Wang, Y., Chen, H., Mahmood, A., Zaib, M., Zhang, W. E., & Sheng, Q. Z. (2022). Towards employing native information in citation function classification. *Scientometrics*, *127*, 6557–6577.

Zhang, Z., Zhang, Y., Sheng, Q. Z., Mahmood, A., Feng, Y., Wang, X., & Zhou, Y. (2024). Multimodal archival data ecosystems. In *2024 IEEE international conference on web services (ICWS), IEEE* (pp. 73–83).

Zhao, H., Luo, Z., Feng, C., Zheng, A., & Liu, X. (2019). A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5206–5215).

Zhao, Y., Tian, Z., Yao, H., Zheng, Y., Lee, D., Song, Y., Sun, J., & Zhang, N. (2022). Improving meta-learning for low-resource text classification and generation via memory imitation. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 583–595).

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

Zhu, Y., Zhou, X., Qiang, J., Li, Y., Yuan, Y., & Wu, X. (2022). Prompt-learning for short text classification. Preprint, arXiv:2202.11345.