Scientometrics (2022) 127:4569–4608
https://doi.org/10.1007/s11192-022-04471-x

SDCF: semi-automatically structured dataset of citation functions

Setio Basuki[1] · **Masatoshi Tsuchiya**[1]

Received: 30 August 2021 / Accepted: 8 July 2022 / Published online: 21 July 2022

Abstract
There is increasing research interest in the automatic detection of citation functions, which is why authors of academic papers cite previous works. A machine learning approach for such a task requires a large dataset consisting of varied labels of citation functions. However, existing datasets contain a few instances and a limited number of labels. Furthermore, most labels have been built using narrow research fields. Addressing these issues, this paper proposes a semiautomatic approach to develop a large dataset of citation functions based on two types of datasets. The first type contains 5668 manually labeled instances to develop a new labeling scheme of citation functions, and the second type is the final dataset that is built automatically. Our labeling scheme covers papers from various areas of computer science, resulting in five *coarse* labels and 21 *fine-grained* labels. To validate the scheme, two annotators were employed for annotation experiments on 421 instances that produced Cohen's Kappa values of 0.85 for *coarse* labels and 0.71 for *fine-grained* labels. Following this, we performed two classification stages, i.e., *filtering,* and *fine-grained* to build models using the first dataset. The classification followed several scenarios, including active learning (AL) in a low-resource setting. Our experiments show that Bidirectional Encoder Representations from Transformers (BERT)-based AL achieved 90.29% accuracy, which outperformed other methods in the *filtering* stage. In the *fine-grained* stage, the SciBERT-based AL strategy achieved a competitive 81.15% accuracy, which was slightly lower than the non-AL strategy. These results show that the AL is promising since it requires less than half of the dataset. Considering the number of labels, this paper released the largest dataset consisting of 1,840,815 instances.

Keywords Active learning · Citation function · *Coarse* label · *Fine-grained* label · Semiautomatic

✉ Setio Basuki
setio@is.cs.tut.ac.jp

Masatoshi Tsuchiya
tsuchiya@is.cs.tut.ac.jp

[1] Department of Computer Science and Engineering, Toyohashi University of Technology, 1–1 Hibarigaoka, Tempakucho, Toyohashi, Aichi 441-8580, Japan

Introduction

Citation analysis is part of the bibliographic analysis that studies how the connection between academic publications is established in terms of one which cites and the other which is cited (Nicolaisen, 2008). Citation analysis has become widespread practice to measure the impact of academic publication. Hlavcheva and Kanishcheva (2020) stated that an academic publication's impact comes from several directions, such as the impact of the researcher, the impact of the group or institution, the local or global academic ranking, and the quality of the publication, which are measured by citation counts. In this setting, the citation counts involve calculating the number of times a document is cited by other documents and is performed through bibliometric databases. However, there is no single database that gathers all publications together with their cited references. The analysis needs to look at several database options, such as Web of Science (WOS),[1] Scopus,[2] Google Scholar,[3] etc. There are several measurements, e.g., h-index personal metric, or impact factor for journal metric, which are widely used as impact indicators because of the citation analysis.

Besides the benefit of current citation analysis, measuring the publication impact using the citation counts gets intense criticism. This is because the citation counts assume that all citations have an equal impact on the academic publication. In fact, not all citations are equal and should not be treated equally (Valenzuela et al., 2015). Treating the citations to be always a positive endorsement of the cited references is problematic because the citations are often made to show disapproval of the cited references. Moreover, the citation analysis fails to capture contextual information (Hirsch, 2005; Mercer et al., 2014) containing several *citation functions*, such as giving the background, using the work, making the comparison, criticism, etc. Focusing on the research paper, the contextual information can be used to dig deeper into the paper. Authors of research papers use citations to show the position of their research in broad literature (Lin & Sui, 2020). The *citation functions* indicate the research's novelty (Tahamtan & Bornmann, 2019), and the quality of the research (Raamkumar et al., 2016), and help authors understand the big picture of the given topics (Qayyum & Afzal, 2018). Furthermore, the *citation functions* enable the research paper to obtain a higher impact when it is used, approved, and supported by other works, and less impact when other works just mention the research paper. Thus, involving the *citation functions* as the contextual information needs serious attention to enrich the impact analysis of the scientific publication.

There is a growing concern for works on the automatic identification of *citation functions* (Pride & Knoth, 2020). This trend is caused by the fact that authors provide citations to determine the *important* and *non-important* roles of citations (Nazir et al., 2020). According to (Zhu et al., 2015), previous works are considered *influential* if they inspire authors to propose solutions. While *incidental* citations refer to a previous work that does not provide a significant impact on the proposed research. In this domain, the terms *important* and *non-important* (Valenzuela et al., 2015) are identical to the terms *influential* and *incidental* (Pride & Knoth, 2020). However, most previous works have a small number of citation instances or considered few types of labels. In addition, existing works have

[1] https://clarivate.com/webofsciencegroup/solutions/web-of-science/.
[2] https://www.scopus.com/home.uri.
[3] https://scholar.google.com/.

suffered from a lack of research variety. Most of these works were developed based on natural language processing (NLP)-related papers. Consequently, several potential *citation functions* were missing from being identified.

The contribution of this paper consists of two parts. In the first part, we propose a new annotation scheme for *citation functions* that have not been accommodated in previous works. Our proposed scheme covers all computer science (CS) fields on arXiv from the beginning to December 31, 2017. This paper uses well-organized parsed sentences of research papers from (Färber et al., 2018) and selects 1.8 million raw *citing sentences*. Based on 5668 randomly selected instances, we developed the proposed annotation scheme following three stages, i.e., top-down analysis, bottom-up analysis, and annotation experiment. Completing the first two stages reveals that there are potential newly proposed labels. We found five *fine-grained* labels related to the *background's* role of *cited papers* that were not proposed by existing works. These labels are *definition*, *suggest*, *technical*, *judgment*, and *trend*. In addition, we found three new labels defining the role of a *cited paper*, i.e., *cited_paper_propose*, *cited_paper_result*, and *cited_paper_dominant*. Our final scheme consists of 5 *coarse* and 21 *fine-grained* labels. Following this, annotation experiments were conducted involving two annotators on 421 samples. We use Cohen's Kappa (Cohen, 1960) to validate the results of the annotation experiments.

The second part of our contribution is to build a dataset of *citation functions* through a semiautomatic approach. This approach was chosen because manual labeling is time-consuming and needs enormous human effort. The proposed method consists of two development stages. In the first stage, we build two classification tasks, i.e., *filtering,* and *fine-grained* classification. The *filtering* task eliminates nonessential labels, and the *fine-grained* task categorizes the detail of the essential labels. In both tasks, we implement a classical machine learning and deep learning approach. Because of the small number of manually labeled instances, pre-trained word embedding methods should be considered here. In addition, this paper demonstrates pool-based active learning (AL) as a low-resource scenario. Following this, the next stage is to assign labels to the entire unlabeled instances using the best models from both tasks of the previous stage.

At the end of this research, this paper delivers several contributions:

- The annotation scheme for citation functions consists of five coarse and 21 fine-grained labels.
- The validity of the scheme is demonstrated in terms of Cohen's Kappa results with 0.85 (almost perfect) for coarse labels and 0.71 (substantial agreement) for fine-grained labels.
- The low-resource scenario-based AL achieves competitive accuracies on less than half of the training data.
- While Bidirectional Encoder Representations from Transformers (BERT)-based AL outperformed other methods in the filtering task, SciBERT reached competitive performances compared to non-AL methods in the fine-grained stage.
- Considering the number of labels, we released the largest dataset consisting of 1,840,815 instances.[4]

[4] https://github.com/tutcsis/SDCF.

The rest of this paper is organized as follows. The "Related works" section describes existing works covering three parts, namely, the annotation schemes of citation functions, the research papers' argumentative structure, and the detection of citation functions. Next, the section "Building the dataset of citation functions" discusses how our dataset is developed. This section covers several points, i.e., scheme development, scheme comparison, annotation strategy, and text classification strategy. The section "Experiment results" presents annotation and text classification experiments, including the released dataset. Finally, in the "Conclusion and future work" section, we present other notable findings from the conducted experiments.

Related works

This section contains a review of existing works related to several points, i.e., the annotation schemes of *citation functions*, the argumentative structures of scientific papers, the dataset of *citation functions*, and the automatic identification of *citation functions*. For consistency, this paper uses several terminologies, namely, *citing paper* is an author's work, *cited paper* is previous work cited by *citing paper*, *citing sentence* is a sentence containing citation marks, and *citation function* is a reason behind citations.

Citation function labels

The review was conducted on previous works proposing their annotation schemes. During the review, we found two major categories of *citation functions*, i.e., *coarse label* (general) and *fine-grained* label (detail). While several works provided both categories, other works provided a single category, either *coarse* or *fine-grained* label. The existing annotation schemes of *citation functions* are shown in Table 1.

We report several notable results while reviewing previous works on *citation functions*. The review of existing works poses the fact that most of the schemes were developed using NLP-related papers. The paper data sources were dominated by ACL Anthology, but several works used other sources such as NIPS Proceedings, PubMed, SciCite, and Computation and Language E-Print Archive. However, we identified two works that have developed the scheme based on multi-disciplinary research papers. In addition, instead of proposing new annotation schemes of *citation functions*, several works reproduced existing schemes. Turning to the developed scheme, most existing works have *citation functions* related to the *background* label, *use*-related labels, and *comparison*-related labels.

Reviewing the labeling scheme of *citation functions* in the previous works reveals several drawbacks.

- Most existing works have developed a few types of labels and the labels were considered too generic. There was a work by (Casey et al., 2019) that proposed detailed labels. However, these labels were designed not only for *citing sentences* but also for other sentences in the Related Work section. This situation brings a consequence that several potential *citation functions* are missing from being identified.
- The labels developed in the previous works were domain-specific since they were created based on Natural Language Processing (NLP)-related papers. As a result, there is an issue related to the compatibility of the labels when applied to broader computer science domains. Here, we identified two works that developed the labels based on multi-

Table 1 Existing works on annotation schemes of citation functions

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
1	Teufel et al. (2006)	Weakness	Weak	Computation and language E-Print archive
		Contrast	CoCoGM	
		Contrast	CoCo-	
		Contrast	CoCoR0	
		Contrast	CoCoXY	
		Agreement/usage	PBas	
		Agreement/usage	PUse	
		Agreement/usage	PModi	
		Agreement/usage	PMot	
		Agreement/usage	PSim	
		Agreement/usage	Psup	
		Neutral	neutral	
2	Dong and Schäfer (2011)	Background	Background	ACL anthology
		Fundamental idea	Fundamental idea	
		Technical basis	Technical basis	
		Comparison	Comparison	

Table 1 (continued)

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
3	Li et al. (2013)	Positive	Based-on	PubMed
		Positive	Corroboration	
		Positive	Discover	
		Positive	Positive	
		Positive	Practical	
		Positive	Significant	
		Positive	Standard	
		Positive	Supply	
		Neutral	Contrast	
		Neutral	Co-citation	
		Neutral	Neutral	
		Negative	Negative	
4	Valenzuela et al. (2015)	Incidental	Related work	ACL Anthology
		Incidental	Comparison	
		Important	Using the work	
		Important	Extending the work	

Table 1 (continued)

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
5	Hernández-Álvarez et al. (2016)	background	Acknowledge	ACL anthology
		background	Corroboration	
		background	Debate	
		Use	Based-on	
		Use	Supply	
		Use	Useful	
		comparison	Contrast	
		Critique	Weakness	
		Critique	Hedges	
6	Jurgens et al. (2018)	Background	Background	ACL anthology
		Motivation	Motivation	
		Uses	Uses	
		Extension/continuation	Extension/continuation	
		Comparison/contrast	Comparison/contrast	
		Future	Future	
7	Bakhti et al. (2018)	Useful	Useful	ACL anthology
		Contrast	Contrast	
		Mathematical	Mathematical	
		Correct	Correct	
		Neutral	Neutral	

Table 1 (continued)

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
8	Casey et al. (2019)	Background	BG-DESC-NE	ACL anthology (related works sections)
		Background	BG-DESC-EP	
		Background	BG-EVAL-P	
		Cited work	CW-DESC	
		Cited work	CW-COMP	
		Cited work	CW-EVAL-P	
		Cited work	A-CW-BUILD	
		Cited work	A-CW-SIM	
		Gap	CW-EVAL-SC	
		Gap	BG-EVAL-SC	
		Author contribution	A-DIFF	
		Author contribution	A-DESC	
		Author contribution	A-GAP	
		Author contribution	A-CW-DIFF	
		Additional labels	OTHER	
		Additional labels	OCR	
		Additional labels	TEXT	

Table 1 (continued)

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
9	Rachman et al. (2019)	Problem	Problem	ACL anthology
		Use	UseModel	
		Use	UseTool	
		Use	UseData	
		Other	Other	
10	Su et al. (2019)	Weakness	Weakness	ACL anthology
		Compare and Contrast	Compare and Contrast	
		Positive	Positive	
		Neutral	Neutral	
11	Zhao et al. (2019)	Use	Use	ACL anthology, NIPS, and PubMed
		Produce	Produce	
		Introduce	Introduce	
		Compare	Compare	
		Extent	Extent	
		Other	Other	
12	Cohan et al. (2019)	Background	Background	SciCite and ACL Anthology
		Method	Method	
		Result comparison	Result comparison	
13	Tuarob et al. (2019)	Utilize	Use	Multiple disciplines
		Utilize	Extend	
		Not utilize	Mention	
		Not utilize	Notalgo	

Table 1 (continued)

No.	Research paper	Coarse classes	Fine-grained classes	Data source domain
14	Pride and Knoth (2020)	Background	background	Multiple disciplines
		Use	Use	
		Compare_contrast	Similarities	
		Compare_contrast	Differences	
		Compare_contrast	Disagreement	
		Motivation	Motivation	
		Extension	Extension	
		Future	Future	

disciplinary fields (Pride & Knoth, 2020; Tuarob et al., 2019), but these works have few and too generic scopes of 8 labels, and 4 labels, respectively. In addition, it is difficult to justify the accuracy of developed labels for comprehensively analyzing the research paper when it is developed according to a wide-ranging domain, for example involving computer science and non-computer science domains. This is because each domain has its style of argumentative structure in the research papers.

To handle these issues, this paper proposed a new labeling scheme of *citation functions* from multiple fields in the computer science domain. Accommodating the variety of *citing sentences* in the multi-field paper and maintaining the scope still in the computer science domain, it is arguable that our proposed labels provide more comprehensive coverage for future *citation function*-related analysis tasks.

Research paper argumentative structure

The argumentative structure represents how information is presented, discussed, and motivated. This structure is useful to justify the scientific claim, state the existing trend, and guarantee research reproducibility (Alliheedi et al., 2019). It is worth discussing argumentative structures in this paper since our proposed annotation scheme naturally contains argumentative labels.

Argumentative structures can be applied to a section-level or sentence-level category. Sollaci and Pereira, (2004) presented the study about the adoption of section-level categories, namely, *introduction*, *methods*, *results*, and *discussion* (IMRAD). This scheme was first used in the 1940s, and since the 1980s, it became the only pattern adopted in health papers. The IMRAD scheme is considered a generic scheme since authors use it to structure a paper's sections. Teufel et al. (1999) developed the first version of *Argumentative Zone* (AZ-I) as a sentence-level category. AZ-I consists of seven labels based on 48 computational linguistics papers. Then, AZ-I was upgraded using 30 Chemistry papers and 9 Computational Linguistics papers (Teufel et al., 2009). The upgraded version, AZ-II, contains 15 labels. The next sentence-level category is *Core Scientific Concepts (CoreSCs)* proposed by (Liakata, 2010). This structure consists of 18 labels based on 265 Physical Chemistry and Biochemistry papers. Another argumentative structure is *Dr. Inventor* proposed by (Fisas et al., 2015). This scheme contains five categories and three sub-categories built based on 40 Computer Graphics papers.

Citation function dataset

Table 2 shows the summary of the existing datasets of citation functions together with estimation number of sample papers and number of labeled instances. The work by (Roman et al., 2021) has provided the largest dataset, consisting of 10 million instances which was labeled automatically. However, these works provided too few labels, i.e., *background, method,* and *result,* which are not sufficient to represent the reason behind citations.

Table 2 Existing datasets of citation functions, togather with the estimation number of source papers, and *citing sentences*

No.	Research paper	Estimation Number of Papers	Estimation number of labeled instances
1	Teufel et al. (2006)	300	9576
2	Dong and Schäfer (2011)	122	1768
3	Li et al. (2013)	91	6355
4	Valenzuela et al. (2015)	20,527	465
5	Hernández-Álvarez et al. (2016)	85	2092
6	Jurgens et al. (2018)	185	1969
7	Bakhti et al. (2018)	?	8700
8	Casey et al. (2019)	95 related work sections	1806
9	Rachman et al. (2019)	Dataset 1: 160 Dataset 2: 50	Dataset 1: 2475 Dataset 2: 1153
10	Su et al. (2019)	?	1432
11	Zhao et al. (2019)	39,601	3088
12	Cohan et al. (2019)	6627	11,020
13	Tuarob et al. (2019)	8063	8796
14	Pride & Knoth (2020)	883	11,233
15	Roman et al. (2021)	?	10 million

Citation function classification

The existing works which performed citation function classification can be divided into two main categories. First, the works that proposed both labeling schemes of citation functions and datasets, second, the works that use other dataset and perform the citation functions classification.

In the first category, the work by (Teufel et al., 2006) is considered as a pioneer in citation functions development. Next, (Valenzuela et al., 2015) built a classification system using support vector machine (SVM) and random forest (RF). Similarly, the RF approach was implemented by (Jurgens et al., 2018) using several features, i.e., pattern, topic, and prototypical. Zhao et al. (2019) used long short-term memory (LSTM), along with character-based embedding, to classify citation resources (tools, code, media, etc.) and functions. Tuarob et al. (2019) proposed a system to classify algorithm citation functions on four usage labels, i.e., *use*, *extend*, *mention*, and *notalgo*. The maximum entropy-based classification was used by (Li et al., 2013) to propose coarse annotation with sentiment labels. Because of the limitation of labeled instances, Dong and Schäfer (2011) introduced ensemble-style self-training to reduce annotation efforts.

Still, in the same category, another work proposing both annotation schemes of citation functions and datasets is (Hernández-Álvarez et al., 2016). This research covered three classification tasks, i.e., citation functions, citation polarities, and citation aspects. All tasks were implemented using sequential minimal optimization. Su et al. (2019) used a convolutional neural network (CNN) for citation function and provenance classification. This task was implemented using multitask learning. Sharing a similar multitask setting. While Bakhti et al. (2018) also used a CNN, Cohan et al. (2019) proposed another multitask learning approach.

In the second category, most of the existing works were dominated by studies focusing on classification strategies based on Valenzuela's dataset. Hassan et al. (2017) proposed six new features combined with Valenzuela's most important features. This work used five algorithms, i.e., SVM, naive Bayes, decision tree, K-nearest neighbor (KNN), and RF. This work outperformed Valenzuela's performance using RF, achieving 84% accuracy. Another work, i.e., Hassan et al. (2018), reached 92.5% accuracy by implementing LSTM using 64 features. Following this, Nazir et al. (2020) proposed using citation frequencies, similarity scores, and citation count. The classification in this research was built using kernel logistic regression, SVM, and RF. Pride and Knoth (2017) used influential and non-influential citations to find highly predictive features. The classification in this work was performed using RF. Next, Wang et al. (2020) used syntactic and contextual features for important and non-important citation detection. This work applied several algorithms, namely, SVM, KNN, and RF.

Besides all these works, Rachman et al. (2019) used a dataset from (Teufel et al., 2006) with re-annotation and developed a model using SVM. Following this, (Roman et al., 2021) used the citation context dataset from CORE. This research applied BERT, depending on the three labels proposed by Sci-Cite (Cohan et al., 2019).

Building the dataset of citation functions

This section describes how our dataset is developed using a semiautomatic approach. The entire system consists of three stages. *The first stage* is annotation scheme development. In this stage, we identified and reviewed the existing labels of *citation functions*. More potential labels are obtained by enlarging the research scopes. The goal of this stage is to develop a final version of the annotation scheme for *citation functions*. *The second stage* is building classification models based on available handcrafted instances. This paper uses several classification scenarios to build these models. The first scenario is implemented using a classical deep learning method. Next, we apply non-contextual and contextual word embedding to cope with limited available data. Furthermore, a low-resource scenario is applied using an AL approach. Finally, *the third stage* is assigning labels to all instances using the best models resulting from the previous stage. Figure 1 depicts how our proposed dataset is developed.

Annotation scheme development

The proposed annotation scheme for *citation functions* in this paper is developed by following several steps. First, we performed top-down and bottom-up analyses. The top-down analysis elaborates on the label definitions of existing schemes, i.e., *background*, *usage*, and *comparison*. In this analysis, the concept of *background* can be expanded by questioning *what*, *why*, *when*, and *how*. The usage can be expanded by categorizing its degree into *inspired*, *uses method*, or *use data*. The comparison can be elaborated using the similarity and difference between *citing paper* and *cited paper*. The bottom-up analysis is used to identify the *citing sentence* patterns in 5,668 random instances. This paper uses a dataset from well-parsed sentences from arXiv (Färber et al., 2018). We filtered sentences containing *<DBLP:*, *<GC:*, *or <ARXIV:* as targeted *citing sentences*. This process results in 1,840,815 *citing sentences* of 15,534,328 sentences. The final scheme consists of 5 *coarse* labels and 21 *fine-grained* labels shown in Table 3.

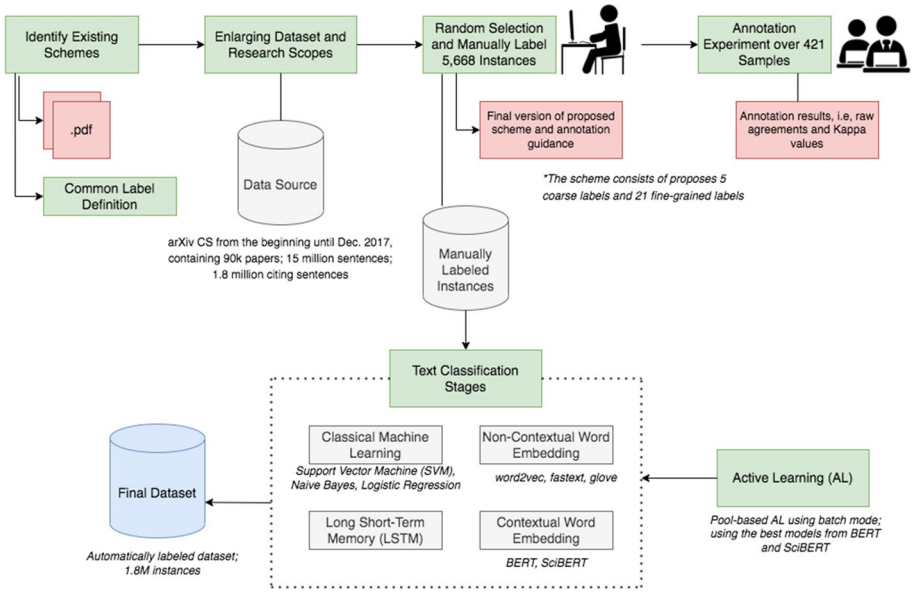

Fig. 1 Development of the semiautomatic dataset of citation functions

Citation scheme comparison

As part of scheme development, a label comparison is performed between our scheme and existing schemes. As mentioned before, the existing schemes consist of *citation functions* and argumentative structures. Through comparison, we show the compatibility and contribution of our proposed scheme. In Table 4, N/A marks indicate the newly proposed labels of our scheme that were not accommodated in existing works. The comparison reveals that our labels are partially and fully compatible with existing labels. However, there exist incompatibilities here. This is caused by the fact that argumentative labels are not naturally designed for *citing sentences.* For example, the label AIM in (Teufel et al., 1999) and (Teufel et al., 2009) is defined as a specific research goal or hypothesis of research papers. This label is commonly stated using ordinary sentences. Another example is the label Conclusion in (Liakata, 2010). This label makes a connection between the experimental results and research hypotheses. Sentences explaining this label naturally are not *citing sentences*. Furthermore, another reason for incompatibility is that labels in the argumentative structure can be represented using more than one sentence.

Annotation strategy

Annotation experiments are the last part of scheme development. Two CS master's degree graduates (annotators) are used in the experiments. Several required resources for the experiments are annotation guidance and unlabeled *citing sentence* samples. In the guidance, there are annotation task explanations, label definitions, and annotation examples, as well as the guidance step-by-step annotation process, best practices, and annotation schedules. After training, each annotator was provided with an Excel sheet containing 421

Table 3 The proposed annotation scheme for citation functions in this paper

Coarse labels	Fine-grained labels	Example of citing sentences
Background: describing the *citing sentences* referring to the theory, principle, concept, topic, problem, etc. from *cited papers*	*Definition:* explaining the definition of general theory, principle, concept, topic, problem, etc	Warped gps <citation> are an extension of gps that allows the learning of arbitrary mappings
	Suggest: giving the reader a suggestion to refer, see more detail, and explore other *cited papers*	For more details on these recurrent activation units, we refer the reader to <citation>
	Judgment: highlighting the positive/negative, useful/not-useful, etc. of concept, topic, problem, etc	The n-coalescent has some interesting statistical properties <citation>
	Technical: explaining how a theory, principle, concept, topic, problem, etc. is applied	An initial decoding is performed with a wfst decoder, using the architecture described in <citation>
	Trend: explaining the significance of the research topic, theory, principle, concept, topic, problem, etc	However, this coherence metric is widely used for the cs scenario due to its simplicity <citation>
Citing paper work: research that is proposed by the author	*Citing_paper_corroboration:* while proposing a research topic, *citing paper* cites *cited paper*	In this section, we define the smoothed analog of the worst-case class and the average-case class <citation>
	Citing_paper_based_on: stating that *citing paper* follow, consider, is built based on, inspired by the *cited paper*	To overcome the difficulty, we come up with an idea inspired by <citation>
	Citing_paper_use: citing paper use, implement, employ, or adopt the concept, dataset, technique, etc	For the simulation experiments, we use the conll data <citation> as annotated data for eight languages
	Citing_paper_extend: citing paper extends, adapts, improves, adds, or modifies the *cited* paper's work	in this paper, we extend the results of pauly <citation>
	Citing_paper_dominant: The performance of *citing paper* outperforms *cited paper's* performance	Our prednet model outperforms the model by <citation>
	Citing_paper_future: mentioning the plan of *citing paper*	To alleviate some of these limitations, we hope to explore near-touch sensors in the future <citation>

Table 3 (continued)

Coarse labels	Fine-grained labels	Example of citing sentences
Cited paper work: what is done by *cited papers*	*Cited_paper_propose:* describing the proposed research by *cited paper*	In <citation> the authors propose a model for storing and using infrared images
	Cited_paper_success: highlighting the success of *cited paper*	Recently, li <citation> successfully use cnn on re-id to extract an effective feature representation
	Cited_paper_weakness: highlighting the weakness of *cited paper*	the limitation of <citation> is that the traffic is assumed to be always cross-directional
	Cited_paper_result: describing the result of *cited paper* (neutral)	However, <citation> reported that users could read text easily on a target of approximately 2 to 3 mm
	Cited_paper_dominant: stating the superiority of *cited paper* compared to *citing paper*	For market-1501 dataset, a recent metric learning approach <citation> outperforms ours
Compare and contrast: Compare and contrast is performed between *citing papers* and *cited papers*	*Compare:* describing the similarity between *citing papers* and *cited papers*	The blht algorithm <citation> is closely related to our work
	Contrast: describing the differences between *Citing papers* and *cited papers*	Unlike <citation> that retains cd, we adopted nce as the basic learning strategy
Other: This label is prepared for *citing sentences* that do not match the above criteria	*Other_cited_paper_comparison:* comparison between *cited papers* (whether similarities or differences between them)	Table compares the computational complexity of the proposed method with aog <citation> and ncte <citation>
	Other_multiple_intent: citing sentences have two or more citation marks for different intents	in <citation> , the mtd system is modeled as a game called pladd, based on flipit games <citation>
	Other_other: This label is designed for *citing sentences* that do not meet all label categories described above	c++ in ilog solver <citation> or java in gecode/j <citation>) and even term rewriting <citation>

Table 4 Comparison between our proposed labels of citation functions and existing schemes

Fine-grained classes of our scheme	Citation function-focused existing works		Argumentative-focused existing works	
	Fully related label	Partially related label	Fully related label	Partially related label
Definition	N/A	Dong and Schäfer (2011): background; Jurgens et al. (2018): background; Zhao et al. (2019): introduce; Cohan et al. (2019): background; Pride and Knoth (2020): background; Roman et al. (2021): background	N/A	Teufel et al. (1999): background; Fisas et al. (2015): background; Liakata (2010): background
Suggest	N/A		N/A	
Judgment	N/A		N/A	
Technical	N/A		N/A	
Trend	N/A		N/A	
Citing_paper_corroboration	Hernández-Álvarez et al. (2016): corroboration	N/A	N/A	N/A
Citing_paper_based_on	Pride and Knoth (2020): motivation; Teufel et al. (2006): Pbas; Dong and Schäfer (2011): fundamental idea; Li et al. (2013): based_on	Su et al. (2019): positive, Casey et al. (2019): A-CW-BUILD; Li et al. (2013): corroboration; Li et al. (2013): discover	N/A	Teufel et al. (1999): basis; Teufel et al. (2009): support; Fisas et al. (2015): approach

Table 4 (continued)

Fine-grained classes of our scheme	Citation function-focused existing works		Argumentative-focused existing works	
	Fully related label	Partially related label	Fully related label	Partially related label
Citing_paper_use	Pride and Knoth (2020): use; Tuarob et al. (2019): use, extend; Teufel et al. (2006): puse; Dong and Schäfer (2011): technical basis; Hernández-Álvarez et al. (2016): based-on, supply; Jurgens et al. (2018): uses; Bakhti et al. (2018): useful; Rachman et al. (2019): useModel, useTool, useData; Zhao et al. (2019): use; Cohan et al. (2019): method	Valenzuela et al. (2015): using the work; Su et al. (2019): positive; Casey et al. (2019): A-CW-BUILD	Teufel et al. (2009): Use	Teufel et al. (1999): basis; Fisas et al. (2015): approach
Citing_paper_extend	Pride and Knoth (2020): extension; Teufel et al. (2006): Pmodi; Valenzuela et al. (2015): extending the work; Zhao et al. (2019): extent; Jurgens et al. (2018): extension/continuation		N/A	Fisas et al. (2015): Approach
Citing_paper_dominant	Teufel et al. (2006): CoCo	Su et al. (2019): compare and contrast	Teufel et al. (2009): ANTISUPP; Liakata (2010): method-new-advantage	Teufel et al. (2009): NOV_ADV; Fisas et al. (2015): outcome, outcome-contribution; Liakata (2010): result
Citing_paper_future	Pride and Knoth (2020): future	N/A	Teufel et al. (2009): FUT; Fisas et al. (2015): future work	N/A

Table 4 (continued)

Fine-grained classes of our scheme	Citation function-focused existing works		Argumentative-focused existing works	
	Fully related label	Partially related label	Fully related label	Partially related label
Cited_paper_propose	N/A	Valenzuela et al. (2015): related work; Hernández-Álvarez et al. (2016): acknowledge; Casey et al. (2019): CW-DESC	N/A	Teufel et al. (1999): other; Liakata (2010): method-old
Cited_paper_success	Casey et al. (2019): CW-EVAL-P; Li et al. (2013): positive	N/A	Liakata (2010): method-old-advantage; Teufel et al. (2009): PREV_OWN	Teufel et al. (1999): other; Teufel et al. (2009): OTHR;
Cited_paper_weakness	Teufel et al. (2006): weak; Hernández-Álvarez et al. (2016): weakness, hedges; Su et al. (2019): weakness; Rachman et al. (2019): problem; Li et al. (2013): negative	Roman et al. (2021): result	(Liakata (2010): method-old-disadvantage	Teufel et al. (1999): other; Teufel et al. (2009): GAP_WEAK;
Cited_paper_result	N/A	Roman et al. (2021): result;	N/A	Teufel et al. (1999): other;
Cited_paper_dominant	N/A	N/A	N/A	
compare	Pride and Knoth (2020): similarities; Teufel et al. (2006): Psim; Casey et al. (2019): A-CW-SIM	Dong and Schäfer (2011): comparison, Valenzuela et al. (2015): comparison	N/A	Teufel et al. (1999): contrast; Teufel et al. (2009): CODI
Contrast	Pride and Knoth (2020): contrast; Teufel et al. (2006): CoCoGM, CoCoR0; Hernández-Álvarez et al. (2016): contrast; Bakhti et al. (2018): contrast	Zhao et al. (2019): compare, Cohan et al. (2019): result comparison Su et al. (2019): compare and contrast	N/A	
Other_multiple_intent	N/A	N/A	N/A	N/A

Table 4 (continued)

Fine-grained classes of our scheme	Citation function-focused existing works		Argumentative-focused existing works	
	Fully related label	Partially related label	Fully related label	Partially related label
Other_cited_paper_comparison	Teufel et al. (2006): CoCoXY; Casey et al. (2019): CW-COMP; Li et al. (2013): contrast	Valenzuela et al. (2015): comparison	N/A	N/A
Other_other	Li et al. (2013): neutral Teufel et al. (2006): Neutral	Bakhti et al. (2018): neutral Su et al. (2019): neutral	N/A	N/A

instances to be labeled. We used *Inter-annotator Agreement* and Kappa values (Cohen, 1960) to validate the annotation results. The Kappa is categorized into several ranges: 0.01–0.20 is stated as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81–1.00 as almost perfect (Wang et al., 2019).

Text classification strategy

Text classification strategies contain two stages, i.e., *filtering* stages and *fine-grained* classification. The *filtering* stage eliminates three *fine-grained* instances belonging to the *coarse* label *other*. The *fine-grained* classification is used to categorize 18 detailed labels. These two stages are applied to a dataset containing manually labeled 5668 instances. Here, we evaluate four classification approaches. First, three classical approaches, namely Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes are used as a baseline system. Then, LSTM is our deep learning method. Considering the few numbers of labeled instances, it is worth applying pretrained word embeddings. We implement two contextual models, i.e., BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019), and three non-contextual models, i.e., fasttext (Bojanowski et al., 2017), word2vec (Mikolov et al., 2013), and glove (Pennington et al., 2014). Note that the non-contextual models' implementations are combined with LSTM. The labeled dataset is divided into training, development, and testing with 80%, 10%, and 10% proportions, respectively. Deep learning approaches are implemented with Keras API, whereas BERT and SciBERT are built using the *ktrain* python library. The best learning rates were obtained during the experiment with a range of 1 e^{-5} to 5 e^{-5}, batch sizes of 32 and 64, and dataset balance or imbalance. The best epoch was specified using early stops by keeping the best model based on validation instances. Regarding the imbalance problem, we use *class_weight* parameter to address this issue. This parameter is applied by multiplying the proportion of minority class to make the distribution of all classes relatively balanced and force to assign higher values to the loss function. Figure 2 depicts the distribution of the development dataset for all classification strategies.

Active learning strategy

Active learning (AL) is subfield of Machine Learning which allows the algorithm to choose to the data from which it learns (Settles, 2010). This method is motivated by existing problems faced by machine learning where the huge unlabeled data is easily obtained but the labels are expensive and time-consuming. The AL argues that the algorithm will perform better using less data because the mechanism for asking queries to the oracle (human annotator) to label the unlabeled instances. The implementation of the AL is conducted by using scenarios in which the learner asks queries. Figure 3 shows the pool-based scenario as the most common scenario of the AL. Lewis and Gale, (1994) define the pool-based AL by assuming there is small set of labeled data L and a large pool of unlabeled data U. The instances are selected from the pool by considering several informativeness measures. Technically, the most informative instances will be labeled by the oracle.

The mechanism to select the most informative instances is called query strategy. The most popular and simplest method of query strategy is uncertainty sampling (Lewis &

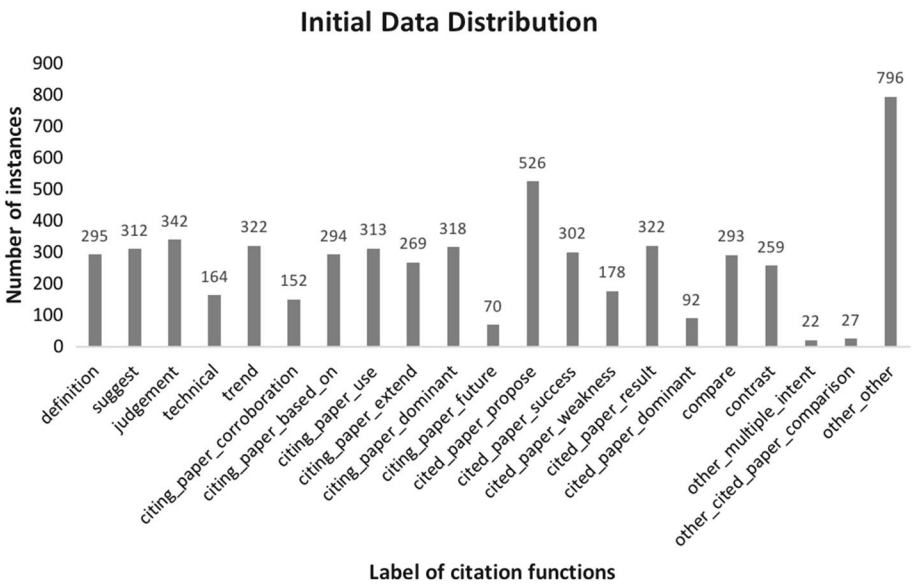

Fig. 2 Development (initial) labeled instance distribution

Gale, 1994) that an instance will be selected when it has the least certain how to label. The uncertainty sampling can be implemented through these sampling variants, by denoting the x_A^* is the most informative instance based on selection method (Settles, 2010):

- *Least confident*
  This is the general uncertainty sampling strategy. Here, the instance will be selected if they have the least confidence in its most likely label. Here, the $\hat{y}$ is the class label having the highest posterior probability of the model θ.

$$x_{LC}^* = \underset{x}{\text{argmax}} 1 - P_\theta(\hat{y}|x)$$

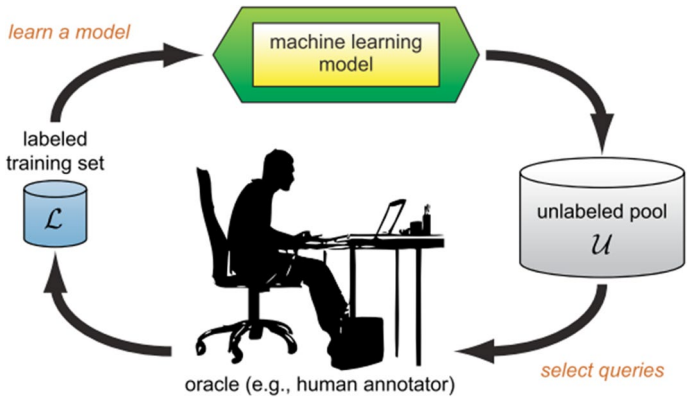

Fig. 3 Pool-based active learning scenario (Settles, 2010)

- *Margin sampling*

  Addressing the drawback of the least confident strategy that considering only the most probable label, the margin sampling selects the instance that has the smallest difference between the most and the second most probable labels. The margin sampling is defined as follows (Scheffer et al., 2001):

$$x_M^* = \underset{x}{\operatorname{argmin}} P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

- *Entropy*

This is the most popular uncertainty sampling strategy that works by utilizing all label probabilities (y_i). Entropy works by using the following formula (Shannon, 1948) to each instance and the instance having the highest value will be queried.

$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

AL has been successfully used to reduce the manual labeling effort. This paper implements the pool-based AL strategy using a batch mode as illustrated in Fig. 4. Using BERT and SciBERT, AL is used in the *filtering* and *fine-grained* stages. The *filtering* stage selects seed L from 10% of training instances, whereas the *fine-grained* stage selects 20% for initial seed L training. The difference in seed proportion is caused by two factors, i.e., the number of available datasets and the number of labels in each stage. The rest of the unlabeled instances U will be used later in AL iterations. The AL strategy is designed to run in 20 iterations. The pretrained word embeddings are trained on seed L. In each iteration, the AL strategy selects a batch consisting of 50 unlabeled instances from U and added them to L with their real labels. This means that there are 1000 new instances from U that are gradually added to L. For batch selection, we compare three sampling approaches, i.e., *least confident*, *max-margin*, and *entropy*. Note that this paper follows the AL strategy proposed by (Ein-Dor et al., 2020; Hu et al.,

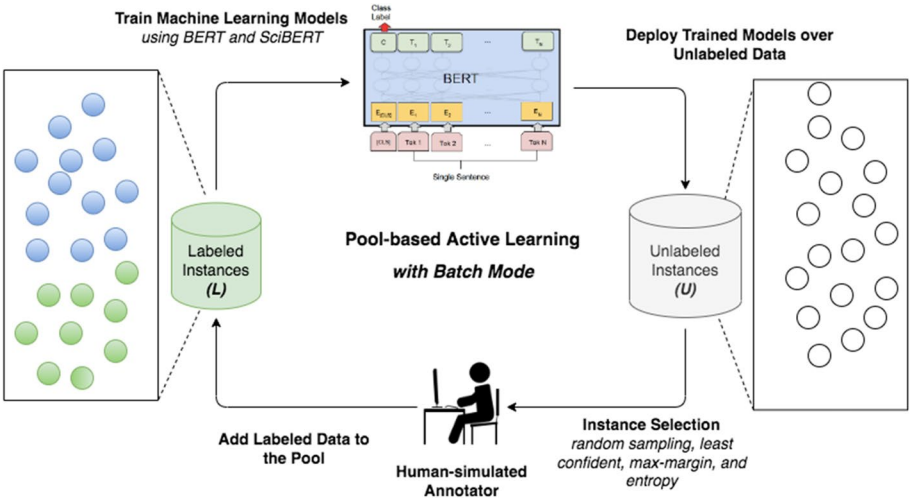

Fig. 4 Pool-based active learning used in this paper, modified from (Settles, 2010)

Table 5 The 2×2 contingency table of the McNemar's test

	Classifier 2: correct	Classifier 2: wrong
Classifier 1: correct	a	b
Classifier 1: wrong	c	d

Table 6 Confusion matrix for inter-annotator agreement on five *coarse* labels

	Background	Citing paper work	Cited paper work	Compare and contrast	Other
Background	**94**	3	2	1	0
Citing paper work	4	**114**	6	0	2
Cited paper work	5	4	**93**	1	0
Compare and contrast	0	2	2	**34**	0
Other	6	3	3	4	**38**

2019) that fine-tuning is performed from scratch in each iteration to prevent overfitting data from previous rounds. The best parameters from a non-AL strategy will be used in the AL experiments.

Statistically significant test

Since this paper implements two classification scenarios, i.e., non-AL and AL, we computed the significance of achieved performances. The McNemar's test (McNemar, 1947) is a statistical test for checking the significance of the difference of paired nominal data. In the case of machine learning, the McNemar's test is used to compare two classifier performances by creating a 2×2 contingency table.

According to Table 5, the test statistic is calculated as follows:

$$X^2 = \frac{(b-c)^2}{(b+c)}$$

Under the null hypothesis where none of the compared classifiers perform better than the other, the test statistic X^2 should be a small value. The high value of X^2 indicate that there is an option to reject the null hypothesis. In addition, we need to specify the common significant threshold by 0.05 and then compute the *p-value*. If the *p-value* is larger than the threshold, then it is called *Fail to Reject Null Hypothesis* which means that none of the compared classifiers perform better than the other. In contrast, if the *p-value* is lower than the threshold, we can *Reject Null Hypothesis* because the two compared classifiers are significantly different. The *p-value* is calculated as follows:

$$p - value = 1 - cdf\left(X^2\right)$$

where *cdf* is cumulative distribution function of the *chi-squared* distribution with 1 degree of freedom.

Table 7 Confusion matrix for Inter-annotator Agreement on *fine-grained* labels

	Definition	Suggest	Judgment	Technical	Trend	Citing_paper_corroboration	Citing_paper_based_on	Citing_paper_use	Citing_paper_extend	Citing_paper_dominant	Citing_paper_future	Cited_paper_propose	Cited_paper_success	Cited_paper_weakness	Cited_paper_result	Cited_paper_dominant	Compare	Contrast	Multiple_intent	Cited_paper_comparison	Other_other
Definition	**16**	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Suggest	0	**21**	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Judgment	1	1	**12**	5	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Technical	2	1	0	**10**	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Trend	1	0	3	0	**13**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Citing_paper_corroboration	0	0	0	0	3	**5**	3	8	1	0	0	1	1	0	2	0	0	0	0	0	0
Citing_paper_based_on	0	0	0	0	0	0	**18**	1	1	0	0	0	0	1	0	0	0	0	0	0	0
Citing_paper_use	0	1	0	0	1	1	1	**17**	0	0	0	0	0	0	0	0	0	0	0	0	1
Citing_paper_extend	0	0	0	0	0	1	0	1	**17**	0	0	0	0	0	0	0	0	0	0	1	0
Citing_paper_dominant	0	0	0	0	0	0	0	0	1	**18**	0	0	0	0	0	0	0	0	0	0	0
Citing_paper_future	0	0	0	0	0	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0

Table 7 (continued)

	Definition	Suggest	Judgment	Technical	Trend	Citing_paper_corroboration	Citing_paper_based_on	Citing_paper_use	Citing_paper_extend	Citing_paper_dominant	Citing_paper_future	Cited_paper_propose	Cited_paper_success	Cited_paper_weakness	Cited_paper_result	Cited_paper_dominant	Compare	Contrast	Multiple_intent	Cited_paper_comparison	Other_other
Cited_paper_propose	0	1	0	1	0	0	0	1	0	0	0	**12**	3	0	3	1	0	0	0	0	0
Cited_paper_success	0	0	0	1	0	1	0	1	0	0	0	1	**13**	1	3	0	0	0	0	0	0
Cited_paper_weakness	0	0	0	0	0	0	0	0	0	0	0	1	1	**16**	0	0	0	1	0	0	0
Cited_paper_result	0	0	0	2	0	0	0	0	0	0	0	2	1	3	**12**	0	0	0	0	0	0
Cited_paper_dominant	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	**19**	0	0	0	0	0
Compare	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**18**	0	0	0	0
Contrast	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	**15**	0	0	0
Multiple_intent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	**15**	2	0
Cited_paper_comparison	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	**17**	0

Table 7 (continued)

	Definition	Suggest	Judgment	Technical	Trend	Citing_paper_corroboration	Citing_paper_based_on	Citing_paper_use	Citing_paper_extend	Citing_paper_dominant	Citing_paper_future	Cited_paper_propose	Cited_paper_success	Cited_paper_weakness	Cited_paper_result	Cited_paper_dominant	Compare	Contrast	Multiple_intent	Cited_paper_comparison	Other_other
Other_other	0	2	1	2	0	0	0	3	0	0	0	1	0	0	2	0	1	0	1	0	**1**

Table 8 The best testing results of each classification technique for the *filtering* stage. Bold values indicate the best result in each performance metric. All metrics are measured by percentage (%)

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
SVM	85.71	82.88	52.28	50.62
Naïve Bayes	85.19	42.59	50.00	46.00
Logistic regression	85.19	70.48	69.67	70.06
LSTM + embedding layer	84.66	50.18	52.62	46.96
LSTM + word2vec	85.19	50.00	42.59	46.00
LSTM + fasttext	85.19	50.00	42.59	46.00
LSTM + glove	85.36	50.60	92.67	47.22
BERT	**90.12**	71.58	**85.15**	75.99
SciBERT	**90.12**	**74.53**	82.72	**77.73**

Experiment results

This section shows the result of the annotation experiments and text classification experiments.

Annotation experiment results

The annotation experiment results contain raw agreement and Kappa values. The confusion matrix in Tables 6 and 7 show raw agreements between annotators. The diagonal bold values in the confusion matrices indicate the number of agreed instances between annotators. The raw agreements reached 88.59% (373 agreed instances) and 72.55% (305 agreed instances) for *coarse* and *fine-grained* labels, respectively. *Citing paper work* achieved the highest percentage of 30.56% in the *coarse* level, followed by *background* with 25.20% and then *cited paper work* with 24.93%. The two labels with the lowest percentage are *other* label with 10.19% and *compare and contrast* label with 9.12%. The *fine-grained* agreements show fairer results since each label has a relatively equal number of samples. The highest percentage in the *fine-grained* level was achieved by *suggest* with 6.89%. Next, *citing_paper_corroboration* and *other* had the two lowest percentages of 1.64% and 0.33%, respectively. The Kappa statistic on *coarse* labels reached 0.85 and 0.71 for the *fine-grained* label. The results are considered as nearly perfect and substantial agreement.

Considering the number of labels in our scheme, the obtained Kappa values are competitive compared with previous works, e.g., (Casey et al., 2019) with 0.77, (Teufel et al., 2006) with 0.72, (Dong & Schäfer, 2011) with 0.757, and (Zhao et al., 2019) with 0.47.

We highlight several sources of disagreement between annotators. The highest number of disagreements in the *coarse* labels occurred in 6 instances where annotator I *(x-axis)* predicted as *background* label and annotator II *(x-axis)* predicted as *other* label. The annotators have an issue to identify the motivation behind the *background* label through its *fine-grained* labels and understanding the motivation behind *other labels*. Focusing on the total of miss-categorized instances by each annotator, there were 15 instances labeled by annotator I and 16 instances labeled by annotator II. On the *fine-grained* labels, the highest disagreement happened on 8 instances where annotator I labeled as *citing_paper_use* and

Table 9 The hyperparameter settings were used in the *filtering* stage

Techniques	Parameters
SVM	ngram_range: (1, 2); imbalance; TF/IDF; kernel = linear
Naïve Bayes	ngram_range: (1, 2); imbalance; TF/IDF
Logistic regression	C: 1; penalty = l1; ngram_range: (1, 1); imbalance; solver = liblinear
LSTM + embedding layer	optimizer = adam; loss = binary_crossentropy; epoch 5; batch 32; imbalance
LSTM + word2vec	optimizer = adam; loss = binary_crossentropy; epoch 5; batch 32; imbalance
LSTM + glove	optimizer = adam; loss = binary_crossentropy; epoch 7; batch 32; imbalance
LSTM + fasttext	optimizer = adam; loss = binary_crossentropy; epoch 5; batch 32; imbalance
BERT	2 e^{-5}; batch 64; imbalance
SciBERT	3 e^{-5}; batch 32; balance

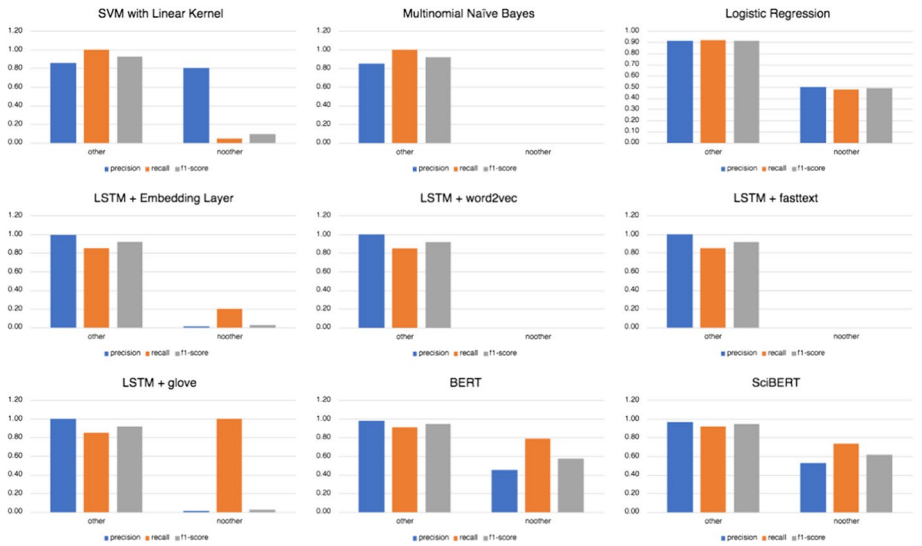

Fig. 5 The performance metrics of individual class in the *filtering* stage. The *x-axis* depicts the classes and their performance metrics, and the *y-axis* depicts the performance values

annotator II labeled as *citing_paper_corroboration*. In this case, both labels are part of the *coarse* label *citing paper work* and our analysis shows that the disagreement on both labels occurred in ambiguous instances. To handle this, the annotation guidelines, including the labeling example, need to be improved to solve the ambiguous instances.

Filtering stage result

In Table 8, we show performance metrics of classification experiments without AL. Focusing on accuracy, the experiments demonstrated that contextual word embeddings, i.e., BERT and SciBERT, shared the highest performances of 90.12%. However, the SciBERT achieved higher macro avg f1 by 77.73% compared with BERT which was

Table 10 The best testing results of each classification technique for *fine-grained* labels

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
SVM	67.29	75.57	66.79	68.49
Naïve Bayes	57.55	74.06	50.94	52.87
Logistic regression	69.98	71.87	70.23	70.53
LSTM + embedding Layer	13.87	10.22	8.09	6.02
LSTM + word2vec	10.97	7.73	2.29	3.45
LSTM + fasttext	14.49	10.02	4.89	5.75
LSTM + glove	14.49	10.23	4.99	6.00
BERT	80.95	80.98	82.40	81.06
SciBERT	**83.64**	**83.46**	**85.35**	**84.07**

Bold values show the best result in each performance metric. All metrics are measured by percentage (%).

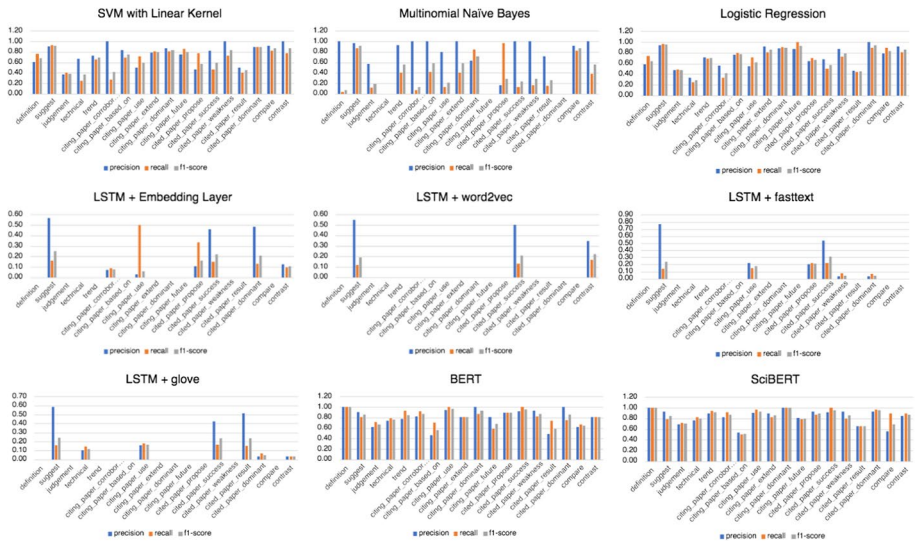

Fig. 6 Performance metrics of each class in the fine-grained stage. The x-axis depicts the classes and their performance metrics, and the y-axis depicts the performance values

only 75.99%. Notably, classical classifiers achieved almost similar accuracies of 85%. But if we look at the macro avg f1, the Logistic Regression reached the highest value by 70.06% among three baseliners. Following this, three non-contextual word embeddings, i.e., word2vec, fasttext, glove, depicted nearly equal accuracies and macro avg precision. But, for macro avg recall and macro avg f1, the glove achieved higher values by 85.15% and 75.99%. Among all methods, the embedding layer showed the poorest performance in all metrics. Table 9 depicts the parameters used in the *filtering* stage.

Looking at the performance of each label in Fig. 5, all performance metrics in the *noother* label are lower than *other* label. There are extreme cases where the *noother*

Table 11 Hyperparameter setting for the best results for *fine-grained* labels

Techniques	Parameters
SVM	ngram_range: (1, 2); TF/IDF; imbalance; kernel = linear;
Naïve Bayes	ngram_range: (1, 2); bag of word; imbalance;
Logistic regression	C: 1; penalty: l1; ngram_range: (1, 2); TF/IDF; imbalance; solver = 'liblinear'
LSTM + embedding layer	epoch 3; batch 32; imbalance; optimizer = adam; loss = categorical_crossentropy;
LSTM + word2vec	epoch 7; batch 32; balance; optimizer = adam; loss = categorical_crossentropy;
LSTM + glove	epoch 7; batch 32; imbalance; optimizer = adam; loss = categorical_crossentropy;
LSTM + fasttext	epoch 7; batch 32; imbalance; optimizer = adam; loss = categorical_crossentropy;
BERT	3 e^{-5}; batch 32; imbalance
SciBERT	3 e^{-5}; batch 32; balance

label has zero values as in Naïve Bayes, word2vec, and fasttext. Two methods, BERT and SciBERT, relatively have balanced proportions compared with other methods.

Fine-grained stage result

As predicted, the performance in this stage will be lower than that in the *filtering* stage. Table 10 shows that there are performance gaps between contextual word embedding and other approaches. The SciBERT showed its superiority compared with other approaches in all metrics. Here, the three non-contextual word embeddings and embedding layers produced the lowest performances below 10% of accuracies and below 10% of macro avg f1. Looking at the baseliners, the best results were achieved by Logistic Regression by around 70% of all metrics. If we look at the individual performance, four approaches i.e., embedding layer, word2vec, fasttext, and glove show poor results (Fig. 6). Here, the three baseline approaches show better performances but still underperform the results from BERT and SciBERT.

All parameter settings in this stage are shown in Table 11. The full performance comparison of BERT and SciBERT in the *filtering* and *fine-grained* stages is shown in Fig. 7.

Active learning results

The experiments were performed using the best parameters from the non-AL results. The filtering experiment used several parameters, i.e., learning rate of 2 e^{-5}, batch size of 64, and imbalanced distribution in the BERT-based AL. For the SciBERT experiments, the best parameters were learning rate of 3 e^{-5}, batch size of 32, and balance distribution. BERT-based *fine-grained* experiment implemented the AL strategies based on learning rate of 3 e^{-5}, batch size of 32, and imbalanced distribution. For SciBERT, the parameters used were learning rate of 3 e^{-5}, batch size of 32, and balanced distribution.

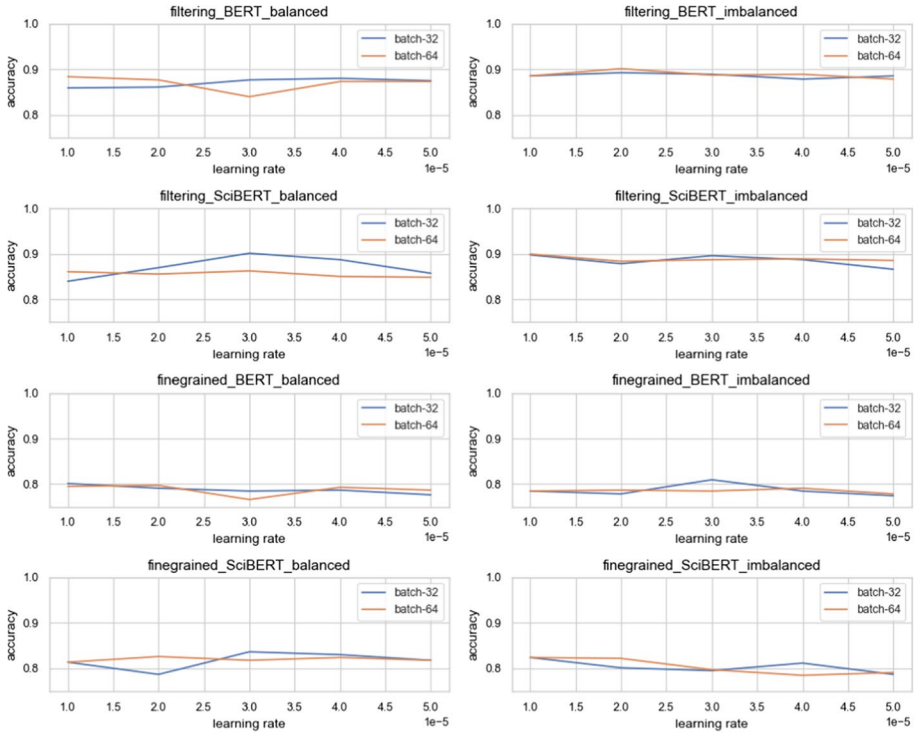

Fig. 7 BERT and SciBERT performance comparison in filtering and fine-grained stages depend on learning rates and batches

Table 12 The best result in the *filtering* stage for AL strategies, and the bold value indicates the highest accuracy among others

Classification strategies	Max_margin		Entropy		Least_confident		Random_sampling	
	Queries	Accuracy (%)	Queries	Accuracy (%)	Queries	Accuracy (%)	Queries	Accuracy (%)
BERT	450	88.71	500	88.88	1000	**90.29**	900	88.35
SciBERT	900	88.00	850	89.59	800	88.88	650	89.41

Filtering stage results

AL-based performance in the *filtering* stage is depicted in Table 12. BERT combined with *least confident* achieved the highest accuracy with 90.29% in the *filtering* stage. To obtain this result, the AL strategy requires 1,000 queried instances for training. While *entropy* used 500 queried instances to obtain 88.88% accuracy, *max-margin* required 450 queried instances to reach 88.71% accuracy. At this stage, the best accuracy reached by SciBERT was 89.59% when integrated with *entropy* on 850 queried instances. Integrating SciBERT with *max-margin* and *least confident* demonstrated the same accuracy of 88.88%, although they need different queried instances, 900 for *max-margin* and

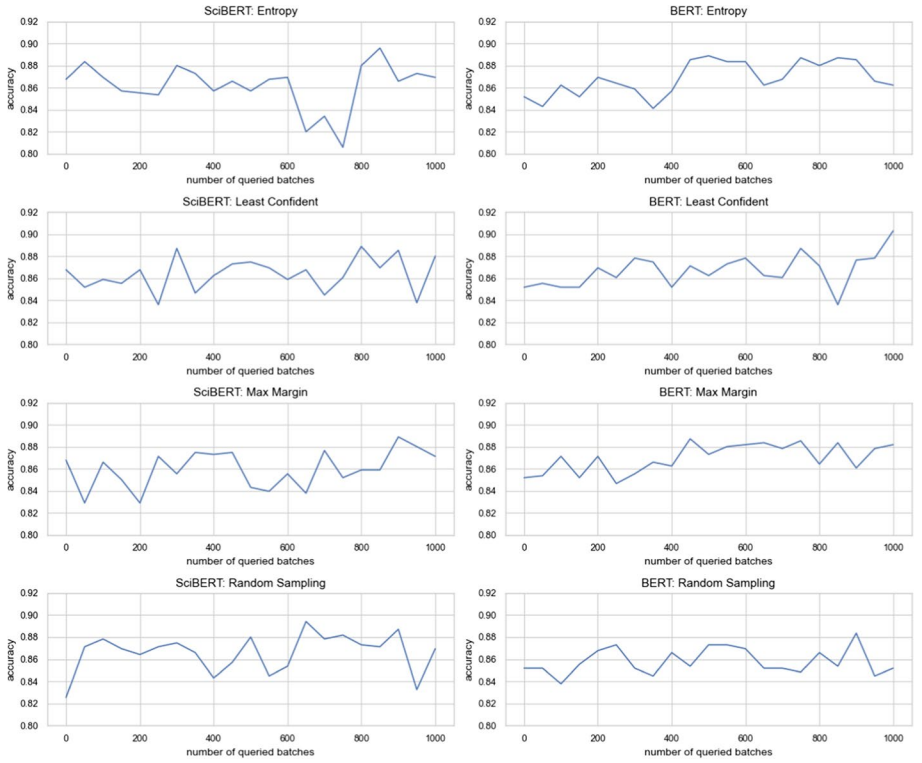

Fig. 8 Result comparison of AL strategies on the *filtering* stage using BERT and SciBERT with four sampling approaches. The data splitting scenario is 1039 (testing), 4534 (simulating L and U), and 453 (seed)

Table 13 Detailed performance metrics of the best accuracy in the AL strategy. All metrics are measured by percentage (%)

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
AL BERT	90.29	77.76	80.19	78.89

800 for *least confident*. The *random sampling* reached the lowest accuracy of 88.35% when the AL was combined with BERT but achieved the second-highest performance by 89.41% in the SciBERT setting. In summary, the AL strategy outperformed the best result from the classification strategy without AL on the entire training instances, especially when integrating BERT with *least confident* and using smaller training instances. The detailed AL results for the *filtering* stage are shown in Fig. 8.

As the AL-based strategy in the filtering stage achieved slightly higher accuracy (90.29%) compared to the non-AL strategy (90.12%), we conducted a statistically

Table 14 The best result of *fine-grained* AL strategies, and the bold value indicates the highest accuracy among others

Classification strategies	Max_margin		Entropy		Least_confident		Random_sampling	
	Queries	Accuracy (%)	Queries	Accuracy (%)	Queries	Accuracy (%)	Queries	Accuracy (%)
BERT	850	79.08	1,000	80.95	700	79.71	650	79.91
SciBERT	850	80.33	850	**81.15**	600	**81.15**	700	80.12

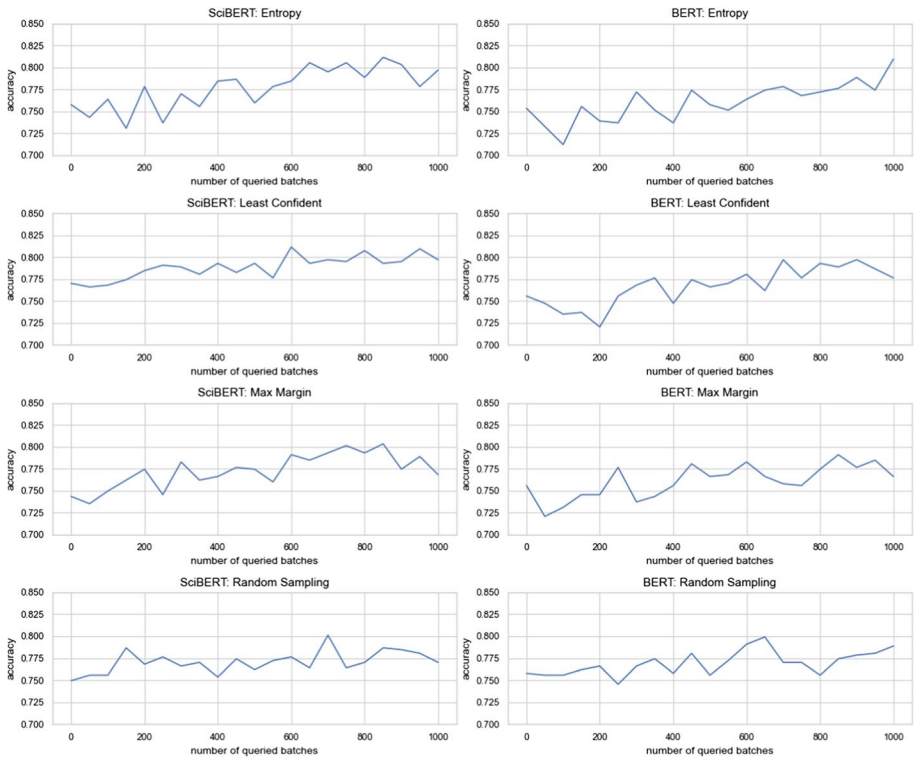

Fig. 9 Result comparison of AL strategies for *fine-grained* classification using BERT and SciBERT with four sampling approaches. The data splitting scenario is 1039 (testing), 3858 (simulating L and U), and 771 (seed)

significant test based on the McNemar approach. Unfortunately, the accuracy achieved using the AL strategy failed to show its significance by producing a *p-value* of 0.73. Instead of relying only on accuracy, we measured alternative metrics as shown in Table 13 as performed in the non-AL setting. Even failed to reject the null hypothesis, we are still able to justify that the AL strategy achieved a better macro avg f1 of 78.89% compared to the best results in the filtering stage by 77.73% using SciBERT.

Table 15 Detailed performance metrics of the best accuracy in the AL strategy. All metrics are measured by percentage (%)

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
AL SciBERT	81.15	81.16	82.83	81.52

Fine-grained results

The AL-based performance in *fine-grained* classification is depicted in Table 14. The highest accuracy of 81.15% was achieved by two AL settings, namely combining SciBERT with entropy-based sampling using 850 queries and combining SciBERT with *least_confident* sampling using 600 instances. Using another sampling technique, i.e., *max-margin*, the AL strategies reached maximum accuracy of 80.33% on 850 queried instances. At this stage, the maximum accuracy obtained by combining BERT and AL was 80.95% on 1000 queried instances. Other sampling methods only reached 79.08%, 79.71%, 79.91% on *max-margin*, *least confident,* and *random sampling*, respectively. The detailed AL results for *fine-grained* classification are shown in Fig. 9.

AL-based strategy in the fine-grained stage achieved slightly lower accuracy (81.15%) compared to the non-AL strategy (83.64%). As in the filtering stage, the significant test conducted in the fine-grained stage compared these two accuracies. The test demonstrated that the accuracy was significantly different with a *p-value* of 0.011. Considering more detailed performances, the AL strategy obtained lower results in all metrics compared to the non-AL strategy, as shown in Table 15.

Here, the AL strategies required fewer instances (less than half of the total dataset) for the training process to achieve competitive accuracy in the *fine-grained* stage and slightly higher accuracy in the filtering stage. This proves two aspects. Firstly, not all instances in the dataset do not share the same contribution toward performance, and secondly, keeping the role of humans in the loop of machine learning using fewer instances will make better judgments than entirely processed datasets by machine learning. Focusing on query strategy, the *least confident* delivered the best performances compared with other methods.

Another point worth mentioning is that the random sampling strategy reached competitive accuracies in the filtering stage when combined with SciBERT and in the fine-grained stage when combined with BERT. In this setting, the random sampling slightly outperformed least confident as the best method in overall scenarios. However, even though it has smallest accuracies compared with all other strategies in another setting, the performances of unbiased instance selection performed by random sampling can be used to generalize the performances when using the whole dataset.

Finally, we use the best models to classify unlabeled *citing sentences*. Table 16 shows the label distribution in the dataset. *cited_paper_propose* has the highest distribution both in the *cited paper work* category and the entire dataset by 243,031 instances, whereas *citing_paper_future* has the lowest instance distribution by 5439. The most interesting point is that there is consistency in the highest distribution in each *coarse* category in the development dataset with manual labeling (See Fig. 2) and the final dataset, e.g., *judgment* for background class, *citing_paper_use* for the *citing paper work* class, *cited_paper_propose* for the *cited paper work* class, and *compare* for the *compare and contrast* class.

Table 16 The distribution of our new dataset of citation function. The bold values indicate the fine-grained labels which have the highest number of instances in each coarse category

Filtering stage	Instance distribution	Coarse label	Fine-grained labels	Instance distribution
No-other	1,328,985	Background	Definition	55,508
			Suggest	51,987
			Judgment	**215,428**
			Technical	85,374
			Trend	66,594
		Citing paper work	Citing_paper_corroboration	113,488
			Citing_paper_based_on	55,878
			Citing_paper_use	**115,215**
			Citing_paper_extend	28,779
			Citing_paper_dominant	24,823
			Citing_paper_future	5,439
		Cited paper work	**Cited_paper_propose**	**243,031**
			Cited_paper_success	34,505
			Cited_paper_weakness	15,054
			Cited_paper_result	154,394
			Cited_paper_dominant	3,215
		Compare and contrast	**Compare**	**39,364**
			Contrast	20,909
Other	511,830	Other	Other	511,830
Total instances				**1,840,815**

Conclusion and future work

This paper developed a dataset of *citation functions* consisting of 1,840,815 labeled instances. The dataset was built using a semiautomatic approach. Specifically, we trained machine learning models on manually labeled data and use these models to label unlabeled instances. Our scheme was developed through top-down analysis, bottom-up analysis, and annotation experiments. Besides our competitive Kappa results, several findings were identified during the experiments. First, assigning *coarse* labels first helped annotators select appropriate *fine-grained* labels. Second, annotation guidance needs to be upgraded to handle ambiguous instances. Third, the proposed scheme is compatible with well-known papers' argumentative structures.

The classification experiments have shown that BERT and SciBERT achieved higher accuracies than other methods. In addition, these two methods achieved promising results using AL on less than half of the training data. SciBERT consistently outperformed BERT in the *fine-grained* stage in both AL and non-AL settings. However, BERT outperformed SciBERT in the *filtering* stage using AL. Note that there is a consistent label distribution between the initial and final datasets.

The limitation of this paper is the labels of *citation functions* are determined using only *citing sentences* themselves, without considering the surrounding sentences. These sentences will be useful during the manual labeling stage, especially when deciding on the labels of ambiguous samples. In future work, we plan to extract sentences before and after the *citing sentences* using the window sizes of two. Not only useful for judging labels of difficult samples, but this information is also important as classification features. Another potential research direction is to investigate the possibilities of applying our scheme of *citation functions* to other research areas through domain adaptation. In this case, domain adaptation becomes a potential method since creating entirely new training data on target domains is expensive, time-consuming, and needs massive human efforts.

Author contribution All authors contributed to the entire research process. The conceptualization, data preparation, methodology, implementation, and manuscript writing were performed by the first author. The second author was responsible for research supervision, study resource preparation, and experimental validation. The final manuscript was approved by all authors.

Funding The first author is a doctorate student at the Department of Computer Science and Engineering at Toyohashi University of Technology (TUT), Japan, which has been granted a scholarship by the Amano Institute of Technology, from 2019 to 2022. The second author is part of the Department of Computer Science and Engineering at TUT. Besides these two institutions, the authors have no support from other institutions.

References

Alliheedi, M., Mercer, R. E., & Cohen, R. (2019). Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining* (pp. 113–123). https://doi.org/10.18653/v1/W19-4514

Bakhti, K., Niu, Z., & Nyamawe, A. (2018). A New Scheme for Citation Classification based on Convolutional Neural Networks. In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE)* (pp. 131–168). https://doi.org/10.18293/SEKE2018-141

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Casey, A., Webber, B., & Glowacka, D. (2019). A Framework for Annotating 'Related Works' to Support Feedback to Novice Writers. In *Proceedings of the 13th Linguistic Annotation Workshop* (pp. 90–99). https://doi.org/10.18653/v1/W19-4011

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 (NAACL)*, (Vol. 1, pp. 3586–3596). https://doi.org/10.18653/v1/N19-1361

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT, 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

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). https://www.aclweb.org/anthology/I11-1070/

Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., & Slonim, N. (2020). Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7949–7962). https://doi.org/10.18653/v1/2020.emnlp-main.638

Färber, M., Thiemann, A., & Jatowt, A. (2018). A high-quality gold standard for citation-based tasks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 1885–1889). https://www.aclweb.org/anthology/L18-1296

Fisas, B., Ronzano, F., & Saggion, H. (2015). On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop* (pp. 42–51). https://doi.org/10.3115/v1/W15-1605

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

Hassan, S.-U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*, 1645–1662. https://doi.org/10.1007/s11192-018-2944-y

Hernández-Álvarez, M., Gómez Soriano, J., & Martínez-Barco, P. (2016). Annotated corpus for citation context analysis. *Latin American Journal of Computing (LAJC), 3*(1), 35–42.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hlavcheva, Y., & Kanishcheva, O. (2020). The reference analysis as a quality characteristic of a scientific article. In *ICTERI-2020 (16th International Conference on ICT in Research, Education and Industrial Applications)* (Vol. 2791, pp. 7–18). http://ceur-ws.org/Vol-2791/2020200007.pdf

Hu, P., Lipton, Z. C., Anandkumar, A., & Ramanan, D. (2019). Active learning with partial feedback. In *The International Conference on Learning Representations (ICLR)* (pp. 1–14). https://arxiv.org/abs/1802.07427

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

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994* (pp. 3–12). https://doi.org/10.1007/978-1-4471-2099-5_1

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*, *September* (pp. 402–407). https://www.aclweb.org/anthology/R13-1052

Liakata, M. (2010). Zones of conceptualisation in scientific papers: a window to negative and speculative statements. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 1–4). https://www.aclweb.org/anthology/W10-3101

Lin, K. L., & Sui, S. X. (2020). Citation functions in the opening phase of research articles: A corpus-based comparative study. *Corpus-Based Approaches to Grammar, Media and Health Discourses*. https://doi.org/10.1007/978-981-15-4771-3_10

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mercer, R. E., Di Marco, C., & Kroon, F. W. (2014). The frequency of hedging cues in citation contexts in scientific writing. In *Advances in Artificial Intelligence 17th Conference of the Canadian Society for Computational Studies of Intelligence* (Vol. 3060, pp. 75–88). https://doi.org/10.1007/978-3-540-24840-8_6

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Conference on Neural Information Processing Systems*. https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Nazir, S., Asif, M., Ahmad, S., Bukhari, F., Afzal, M. T., & Aljuaid, H. (2020). Important citation identification by exploiting content and section-wise in-text citation count. *PLoS ONE, 15*(3), 1–19. https://doi.org/10.1371/journal.pone.0228885

Nicolaisen, J. (2008). Citation analysis of the contact lens field. In *Annual Review of Information Science and Technology* (Vol. 41, Issue 1). https://doi.org/10.1002/aris.2007.1440410120

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Pride, D. & Knoth, P. (2017). Incidental or influential? A decade of using text-mining for citation function classification. In *16th International Society of Scientometrics and Informetrics Conference*. https://doi.org/10.5860/choice.51-2973

Pride, D. & Knoth, P. (2020). An authoritative approach to citation classification. In *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)* (pp. 337–340). https://doi.org/10.1145/3383583.3398617

Qayyum, F., & Afzal, M. T. (2018). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics, 118*, 21–43. https://doi.org/10.1007/s11192-018-2961-x

Raamkumar, A. S., Foo, S., & Pang, N. (2016). Survey on inadequate and omitted citations in manuscripts: A precursory study in identification of tasks for a literature review and manuscript writing assistive system. *Information Research, 21*(4), 733.

Rachman, G. H., Khodra, M. L., & Widyantoro, D. H. (2019). Classification of citation sentence for filtering scientific references. In *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019* (pp. 347–352). https://doi.org/10.1109/ICITISEE48480.2019.9003736.

Roman, M., Shahid, A., Khan, S., Koubaa, A., & Yu, L. (2021). Citation intent classification using word embedding. In *IEEE Access* (Vol. 9, pp. 9982–9995). https://doi.org/10.1109/ACCESS.2021.3050547

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden markov models for information extraction. In *Advances in Intelligent Data Analysis (IDA) 2001. Lecture Notes in Computer Science*, (Vol. 2189, pp. 309–318). https://doi.org/10.1007/3-540-44816-0_31

Settles, B. (2010). Active learning literature survey. *Computer Sciences Technical Report*. https://doi.org/10.1016/j.matlet.2010.11.072

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey. *Journal of the Medical Library Association, 92*(3), 364–367.

Su, X., Prasad, A., Sugiyama, K., & Kan, M. Y. (2019). Neural multi-task learning for citation function and provenance. In *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2019-June* (pp. 394–395). https://doi.org/10.1109/JCDL.2019.00122.

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

Teufel, S., Siddharthan, A., & Batchelor, C. (2009). Towards discipline-independent Argumentative Zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1493–1502). https://www.aclweb.org/anthology/D09-1155

Teufel, S., Carletta, J., & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 110–117). https://www.aclweb.org/anthology/E99-1015

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings Ofthe 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 103–110). https://www.aclweb.org/anthology/W06-1613

Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S.-U., & Haddawy, P. (2019). Automatic classification of algorithm citation functions in scientific literature. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2019.2913376

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Wang, J., Yang, Y., & Xia, B. (2019). A simplified Cohen's kappa for use in binary classification data annotation tasks. *IEEE Access, 7*, 164386–164397. https://doi.org/10.1109/ACCESS.2019.2953104

Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020). Important citation identification by exploiting the syntactic and contextual information of citations. *Scientometrics, 89*, 2109–2129. https://doi.org/10.1007/s11192-020-03677-1

Zhao, H., Luo, Z., Feng, C., & Ye, Y. (2019). A context-based framework for resource citation classification in scientific literatures. In *SIGIR'19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1041–1044). https://doi.org/10.1145/3331184.3331348

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