The Crucial Role of Citation Functions in The Technology-assisted Peer Review

（計算機によって支援されたピアレビューにおける引用目的の役割に関する研究）

January, 2023

Doctor of Philosophy (Engineering)

Setio Basuki

セティオ　バスキ

Toyohashi University of Technology

Abstract

This research aims to develop a citation functions-based prediction method of paper quality to support Technology-assisted Peer Review (TAPR). The prediction method is intended to reduce the review burden which becomes a critical issue in today's paper submission process. Since the review burden problem has gained much attention, many works have developed the TAPR system to handle this issue. However, most of existing works were created by involving reviewers' comments which is considered unapplicable for reducing the review burden. Addressing this issue, this research proposes a prediction method to estimate the paper quality depending only on the paper itself. The estimator of paper quality used in this thesis is citation functions which represent the reason why author of research paper cites previous works. Moreover, the citation functions present the position of the proposed research in wide-ranging literature, understand the broad view of the given research topics, indicates the novelty of the proposed research, and estimate the quality of the proposed research.

The challenge for estimating the paper quality using citation functions will depend on whether the labels are representative enough to capture all potential citation roles in the full text of research paper. Handling drawbacks of current available scheme of citation functions that have a small number of citation instances, few types of labels, and suffer from lack of research variety, this research proposes a new labeling scheme of citation functions covering multi-field computer science domains consisting of 5 coarse labels and 21 fine-grained labels. The annotation experiments on the proposed scheme achieved Cohen's Kappa values of 0.85 for coarse labels and 0.71 for fine-grained labels. The scheme is then used to construct a large dataset of citation functions using a semiautomatic approach which follows two classification stages, i.e., filtering and fine-grained. Adopting the Active Learning (AL) techniques using less than half of training dataset, the Bidirectional Encoder Representations from Transformers (BERT)-based AL in the filtering stage and SciBERT-based AL in the fine-grained stage reached the accuracies of 0.90 and 0.81, respectively. Finally, this research released the largest dataset consisting of 1,840,815 instances.

The prediction method for TAPR, which covers two classification tasks and one regression task, is developed based on the proposed scheme and the best models to create the dataset. While the classification tasks focus on predicting the final review decision (accepted-rejected) and estimating the paper quality (good-poor), the regression task is used to predict the peer review scores. Both classification and regression are implemented using three features i.e., citing sentence features developed based on labeling scheme of citation functions, regular sentence features created by applying the label of citation functions to non-citation text, reference-based features constructed by identifying the source of citations. The classification experiments on the International Conference on Learning Representations (ICLR) 2017-2020 showed that the proposed methods are more effective in the good-poor task compared to the accepted-rejected task by demonstrating the best accuracy of 0.75 and 0.73, respectively. Obtaining as many good papers as possible in the good-poor task, this research reached a satisfying recall of 0.99 by using only the citing sentence features. The regression experiments indicate that the best result in predicting the average review score is higher than in the individual review score by showing RMSE of 1.34 and 1.71, subsequently.

As mentioned above, reaching as high recall as possible is important to get as many good papers as possible, which is more reasonable and applicable for supporting the editor to filter the submitted manuscripts. Interestingly, this highest recall was reached by using only the citing sentences feature. These results prove the hypothesis of this research about the crucial role of citation functions in the manuscript.

Table of Content

List of Figures

List of Tables

List of Abbreviations

AAAI	: Association for the Advancement of Artificial Intelligence
ACL	: Association for Computational Linguistics
ACM	: Association for Computing Machinery
AI	: Artificial Intelligence
AIRA	: The Artificial Intelligence Review Assistant
AISTATS	: The International Conference on Artificial Intelligence and Statistics
AL	: Active Learning
API	: Application Programming Interface
ARRIVE	: Animal Research Reporting of In Vivo Experiments
AUC	: Area under the ROC Curve
AZ	: Argumentative Zoning
BERT	: Bidirectional Encoder Representations from Transformers
CNN	: Convolutional Neural Network
CoNLL	: Conference on Computational Natural Language Learning
CORD-19	: COVID-19 Open Research Dataset
CORE	: Aggregating the world's open access research papers
CoreSC	: The Core Scientific Concepts
COVID-19	: Coronavirus Disease of 2019
CS	: Computer Science
CVPR	: Conference on Computer Vision and Pattern Recognition
DBLP	: Digital Bibliography and Library Project
DL	: Deep Learning
DTR	: Decision Tree Regressor
ECCV	: European Conference on Computer Vision
EMNLP	: Empirical Methods in Natural Language Processing
EVISE	: Editorial System by Elsevier
FN	: False Negative
FP	: False Positive
FS	: Feature Selection
GBR	: Gradient Boosting Regressor
IAA	: Inter-Annotator Agreement
ICASSP	: The International Conference on Acoustics, Speech, and Signal Processing
ICCV	: International Conference on Computer Vision

ICLR	: The International Conference on Learning Representations
ICML	: International Conference on Machine Learning
ICRA	: The International Conference on Robotics and Automation
IEEE	: Institute of Electrical and Electronics Engineers
IJCAI	: The International Joint Conference on Artificial Intelligence
JMLR	: The Journal of Machine Learning Research
KNN	: k-Nearest Neighbors
LSTM	: Long Short-term Memory
MAE	: Mean Absolute Error
MDAR	: The Materials Design Analysis Reporting
MIT Press	: Massachusetts Institute of Technology Press
ML	: Machine Learning
NAACL	: North American Chapter of the Association for Computational Linguistics
Neuralcom	: Neural Computation
NeurIPS	: (formerly NIPS) Conference on Neural Information Processing Systems
NIH	: National Institutes of Health
NISP	: Conference on Neural Information Processing Systems
NLP	: Natural Language Processing
PMC	: PubMed Central
PubMed	: Public/Publisher MEDLINE
RF	: Random Forest
RFE	: Recursive Feature Elimination
RFR	: Random Forest Regressor
RMSE	: Root Mean Square Error
RS	: Rhetorical Structures
SciBERT	: Scientific Bidirectional Encoder Representations from Transformers
SFS	: Sequential Feature Selector
SIGKDD	: Special Interest Group on Knowledge Discovery and Data Mining
SMOTE	: Synthetic Minority Oversampling Technique
SOTA	: State of The Art
STM	: International Association of Scientific Technical and Medical Publishers
SVM	: Support Vector Machine
SVR	: Support Vector Regression
TAPR	: Technology-Assisted Peer Review
TN	: True Negative

TP	: True Positive
XGBR	: Extreme Gradient Boosting Regression

Chapter 1 Introduction

1.1. The Peer Review

Peer Review is one of the most important stages in the scientific publication. The principles that exist in the peer review not only be applied to journal publishing but also in the conference submission and grant proposal assessment (Rowland, 2002). In the traditional peer review system, through Figure 1.1, (Checco et al., 2021) stated that the review process can be summarized as follows. After manuscript submission, the initial checks, e.g., plagiarism detection, manuscript formatting, metadata, etc., will be conducted. These initial checks are often accompanied by checking the argument quality and the relevance of submitted manuscripts to the journal. If the submitted manuscript passes the initial check, then it will be passed to the next stage. Here, the manuscript will be evaluated by reviewers on several indicators, including novelty, originality, significance, soundness, etc. Before the final decision whether accepted or rejected, this process can be iterative since the reviewers often suggest authors for revisions. Regardless of whether the review will be done based on single-blind or double-blind, this process will require a lot of effort.

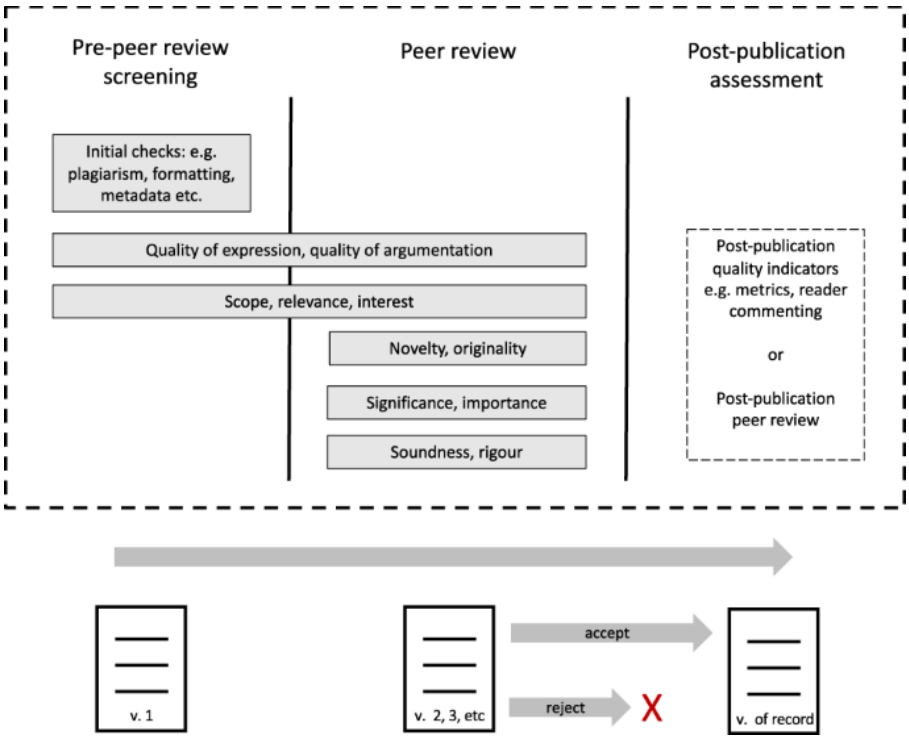


Figure 1.1. The general stages of the peer review process.

The massive number of research papers being published makes the peer review process more challenging. According to the STM Report 2018 (Johnson et al., 2018), there are 33,100 peer-reviewed English-language journals and 9,400 non-English-language journals that publish more than 3 million articles per-year. Another report says that the yearly process in reviewing the manuscripts that were previously rejected reaches 15 million hours (Checco et al., 2021). Changing the perspective to the conference management system, EasyChair, a web application of conference management system, manages more than 100,000 conference events since 2002[1]. The peer review process faces other problematic issues related to the huge review burden which leads to a phenomenon which is commonly called as "reviewer fatigue" (Breuning et al., 2015; Fox et al., 2017). This "fatigue" is caused by situation that each submitted paper will be handled by 2 or 3 reviewers and a handling editor. The study on the reason why potential reviewers unable to review reveals the fact that they are willing to review but they are too busy or obtained too many review invitations (Breuning et al., 2015). The situation going worser when there is an uneven geographical distribution of the experts does the review puts the peer review process into an over-burdened system (Jubb, 2016).

Another problematic issue of peer review process is that the review relies only on human expertise, it will unavoidably tend to be biased and subjective because of several situations such as expert academic background, experience, emotion and health (Tong et al., 2021). Other problematic issues were well identified by (Tennant, 2018) by showing several points, i.e., lack of training on how to do the peer review or response the review (Schroter et al., 2004), relationship between the journal quality and peer review quality (Pierson, 2018), and standard core competencies for editor (Moher et al., 2017). In addition, (Jana, 2019) explained additional drawbacks of traditional peer review system, such as expensive and publication delay, harsh comments caused by reviewers' anonymity, author-recommended reviewers, and irresponsible reviewers to complete the review process. In handling these issues, (Tong et al., 2021) posed an opportunity for involving Artificial Intelligence (AI) in the peer review process, which is stable, innovative, optimal, and efficient. The AI-based Technology-Assisted Peer Review process is not intended to replace the whole human's intervention but more focus on collaboration between the technology and all stakeholders in the review process.

1.2. Technology-Assisted Peer Review

The Technology-assisted Peer Review (TAPR) gains much attention due to the massive burden of globally research publishing. (Tong et al., 2021) divided the TAPR, especially that is developed using AI, into three task categories, i.e., submission review, reviewer recommendation, academic influence prediction. In the submission review tasks, the AI can be used to review the paper structure and to validate the data. Moreover, this task involves the detection of academic misconduct. While the reviewer recommendation task focuses on determining the most suitable reviewers, the paper influence prediction is used to estimate the future influence of submitted papers. The TAPR tools have been developed by both publishers and technology vendors for different purposes. For example, *Frontiers*[1] has developed The Artificial Intelligence Review Assistant (AIRA)[2] which addressed several tasks such as reducing reviewer fatigue, editor-article matching, connecting with funders, etc. The next tool is UNSILO Evaluate Technical Check[3] which evaluates how well the submitted manuscript follows the submission guideline. The *SciScore*[4] offers a service to analyze method section of the paper, based on several standard of reporting such as National Institutes of Health (NIH)[5], The Materials Design Analysis Reporting (MDAR)[6], Animal Research: Reporting of In Vivo Experiments (ARRIVE)[7], etc. and the provides scores for every submission. Following this, *Scholastica*[8] optimize the peer review through integrating the peer review itself with the production and journal hosting software. Elsevier has released editorial tool called EVISE [9] for several tasks including plagiarism detection and reviewer matching (Tennant, 2018). All these developed tools proofs that the peer review system needs to be intervened by technologies to solve the issues of review burden.

[1] https://www.frontiersin.org/
[2] https://blog.frontiersin.org/tag/aira/
[3] https://discovery.researcher.life/publisher
[4] https://sciscore.com/
[5] https://www.nih.gov/
[6] https://www.pnas.org/doi/10.1073/pnas.2103238118
[7] https://arriveguidelines.org/
[8] https://scholasticahq.com/
[9] https://www.elsevier.com/connect/reviewers-update/rolling-out-our-new-editorial-system-evise

1.3. Estimating Paper Quality through Citation Functions

Among these three TAPR tasks, estimating the paper quality needs more attention since it requires iterative stages of reviews. Assessing paper quality not only requires a lot of effort but also gets intense criticism related to the inconsistency among review results, review scores, and final decisions (Kravitz et al., 2010). To estimate the paper quality, **the existence of appropriate *citation functions* is crucial to be assessed during peer review process**. The *citation functions* show the reason why authors of academic papers cite previous works (Teufel et al., 2006). The citations themselves cannot be treated equally, since it indicates different functions (Valenzuela et al., 2015), such as stating background, doing a comparison, writing criticism, etc. The existence of the citation brings several advantages in estimating the paper quality during review process such as identifying the positioning of the proposed research in wide-ranging literature (Lin & Sui, 2020), understanding the broad view of the given research topics of the paper (Qayyum & Afzal, 2018), indicating the novelty of the proposed research (Tahamtan & Bornmann, 2019), and estimating the quality of the proposed research (A. J. Casey et al., 2019; Raamkumar et al., 2016). Following this, the *citation functions* will bring the research impact analysis to the next level, by addressing the drawbacks of existing methods developed based on two principles, namely (a) citation counts-based impact analysis, (b) assuming the citations to be always a positive endorsement. As a result, existing impact analysis failed to capture contextual information (Hirsch, 2005; Mercer et al., 2014). Considering these significant roles of *citation functions*, it can be potential a predictor for judging the research paper's quality.

The challenge to use the *citation functions* to predict the paper quality is providing comprehensive labels to capture all potential citation roles in the full text research paper. Reviewing the literature on developing labeling schemes of *citation functions* reveals several drawbacks. Most previous works have a small number of citation instances or considered few types of labels. There was a work by (A. Casey et al., 2019) that proposed detailed labels. However, these labels were designed not only for *citing sentences* (the sentence containing citation marks) 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. In addition, existing works have suffered from a lack of research variety. Most of these works were developed 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. This thesis identified three works that developed the labels based on multi-disciplinary fields (Pride & Knoth, 2020; Roman et al., 2021; Tuarob et al., 2019), but these works have few and too generic scopes of 4 labels, 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. Another drawback is that the final datasets of *citation functions* generated in existing works were small. The work of (Roman et al., 2021) has 10 million instances and has been developed based on multi-disciplinary fields, but it only contains three labels. Therefore, there is a necessity to provide comprehensive the scheme and dataset of *citation functions* in order to make them as main indicator for predicting the paper quality as technology-assisted peer review.

The development of technology-assisted peer review to predict the quality, final review decision, and review score of research papers gained much attention. Existing works on prediction of paper final acceptance were built using various purposes, ranging from predicting three-classes outcomes accepted, borderline, and rejected (K. Wang & Wan, 2018) to suggesting two-labels outcome accepted and rejected as majority targeted classes as in (Fytas et al., 2021). However, existing works have two main drawbacks. First, due to the criticism in the peer review process related to the inconsistency among review results, review scores, and final decisions (Kravitz et al., 2010), directly predicting final review decision leads to a bias in determining the quality of the paper. Handling these issues, the prediction method in this thesis not only covers the *accepted-rejected* prediction task but also covers the *good-poor* prediction task and review score prediction task which are more reasonable. Second, most of existing works use the review comments as prediction features. This is considered unfair and unapplicable when the main goal is to reduce the review burden in the peer review process. Therefore, the prediction method does not depend on the review comments is necessary.

1.4. Contributions of this Thesis

Doing the literature review on TAPR poses several research gaps which can be divided into two main categories. The first category is that none of existing works in this topic consider the

important roles of *citation functions* for estimating the quality of research papers. Following this, the second category is that most the works use the review comments when predicting the paper quality which is considered unfair and unapplicable when the main issue is to reduce the review burden.

The contribution of this thesis is proposing a method for predicting paper quality using *citation functions* to support TAPR. The approach has been realized through several stages which is summarized as follows:

- Developing **a new labeling scheme of *citation functions*** from multi-field CS domains. Accommodating the variety of *citing sentences* in the multi-field papers and maintaining the scope still in the computer science domain, it is arguable that the proposed labels provide more comprehensive coverage for future *citation function*-related analysis tasks. The labeling scheme is developed by following a top-down (identifying the definition of potential labels) and bottom-up (gathering the potential pattern in the dataset) approach. For validating the proposed scheme, this thesis performs the annotation experiments and measures the inter annotator agreement (IAA) in terms of raw agreement percentage and Cohen's Kappa values.
- **Creating a dataset of *citation functions*** using semi-automatic approach. This approach randomly selects samples of whole extracted *citing sentences* as development dataset and develops machine learning (ML) model based on this dataset. The classification experiments will be conducted using two main scenarios. First, the classification using the whole development dataset combined with several machine learning techniques such as classical ML algorithms, deep learning (DL), and pre-trained word embeddings. Second, the classification experiments using low resources scenarios based on active learning (AL). The AL will take a few samples which are called seed to develop the initial model and then iteratively update the model. The best model is then used to classify the whole dataset.
- Proposing a prediction method to **automatically estimate paper quality based on *citation functions***. The method covers two classification tasks and a regression task. The classification tasks are used to predict the final review decision (*accepted-rejected*) and paper quality (*good-poor*), and the regression task is intended to predict the review score (average score and individual score). The prediction method is realized through developing three types of predictors, i.e., *citing sentence* predictor developed based on

labeling scheme of *citation functions*, *regular sentence* predictor created by applying the label of *citation functions* to non-citation text, *reference-based* predictor constructed by identifying the source of citations. Moreover, this study adds one more prediction called combination predictor by incorporating all mentioned predictors. Both classification and regression tasks will use the same predictors accompanied by several feature selection techniques.

- Building the **dataset of *citation functions* for COVID-19 academic papers**. Because the preparation of new labels of *citation functions* and building a new dataset requires much human effort and is time-consuming, this thesis uses the labeling scheme of *citation functions* that were built for the CS domain. This thesis uses the COVID-19 Open Research Dataset (CORD-19), and the extracted *citing sentences* are automatically categorized using the classification models built from the CS domain.

1.5. Technical Terms

This thesis uses several technical terms for consistency. These terms are used in the entire thesis. The term ***citing paper*** is a paper other works and the ***cited paper*** is a paper cited or mentioned by the *citing paper*. Following this, ***citing sentences*** is used to define the sentence in the paper containing citation marks. The term ***regular sentences*** are the sentences in the paper that do not contain citation marks. Next, the term ***citation functions*** which represents the reason behind citation.

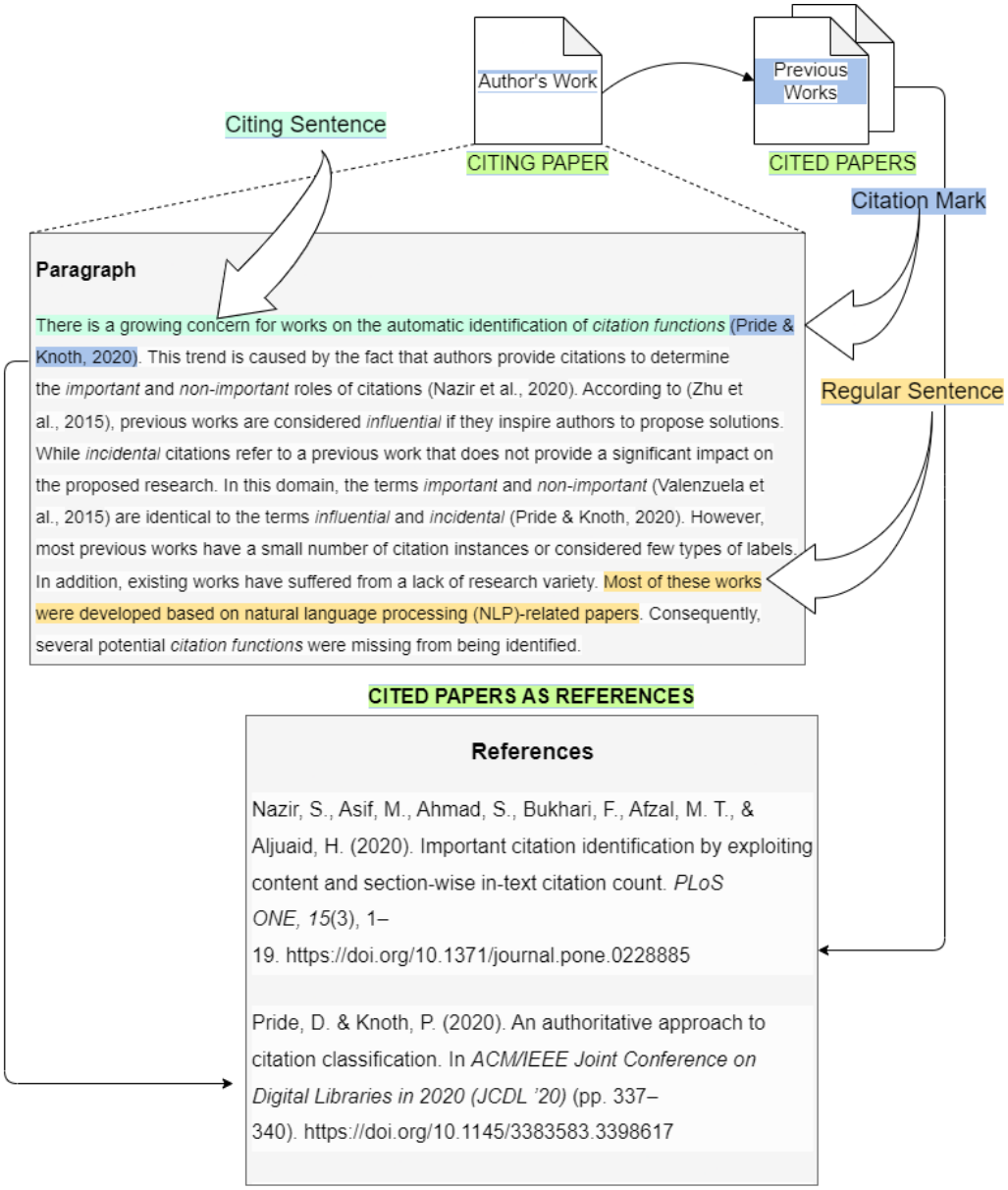


Figure 1.2. Technical terms used in this thesis.

1.6. Outline of The Thesis

This thesis is organized as follows. **Chapter 2** explains the development of a new labeling scheme of *citation functions*. This chapter will cover the existing scheme of *citation functions* and the annotation experiments on the newly proposed scheme. In **Chapter 3**, this thesis describes the development of dataset of *citation functions* using semi-automatic approach. The approach involves several ML techniques such as classical model, deep learning model, word embedding model, and active learning. **Chapter 4** describes the development of a prediction method for estimating paper quality automatically. The prediction model covers three tasks, i.e., final review decision *(accepted-rejected)*, paper quality *(good-poor)*, and review scores

(average score and individual score). **Chapter 5** presents the experiments on applying the scheme of *citation functions* developed based on CS domain to COVID-19 domain. This chapter uses dataset of COVID-19-related research papers.

Chapter 2 – The Development of a New Labeling Scheme of Citation Functions

This chapter explains the new labeling scheme of *citation functions* proposed in this thesis. This proposed scheme is developed to address the drawbacks of existing schemes by enlarging the research scope and number of potential labels. The scheme development presented in this chapter covers several parts. The first part discusses existing schemes of *citation functions* and reveals several limitations on them. The second part presents the argumentative structure of the research paper to explain how information in the paper is organized and its relationship with the *citation functions*. Next, the third part explains how the new labeling scheme of *citation functions* is created by following the top-down and bottom-up approaches. Following this, the fourth part conducts the label comparison between the new scheme proposed in this thesis with existing scheme of *citation functions* and current available paper argumentative structure. The last two parts focus on performing the annotation experiments and the evaluation of the annotation results.

2.1. Existing Labeling Scheme of Citation Functions

The review was conducted on previous works proposing their annotation schemes. During the review, this study 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 2.1.

Table 2.1. Existing works on annotation schemes of citation functions.

No.	Research Paper	Coarse Classes	Fine-Grained Classes	Data Source Domain
1	(Simone 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 & Schäfer, 2011)	Background	Background	ACL Anthology
		Fundamental idea	Fundamental idea	
		Technical basis	Technical basis	
		Comparison	Comparison	
3	(X. 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	
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	
8	(A. 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	
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	
14	(David Pride & Knoth, 2020)	background	background	Multiple Disciplines
		use	Use	
		compare_contrast	Similarities	
		compare_contrast	Differences	
		compare_contrast	Disagreement	
		motivation	Motivation	
		extension	Extension	
		future	Future	

This study reports 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, this thesis 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 (A. 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, there are two works that developed the labels based on multi-disciplinary fields (David 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 thesis 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 the proposed labels provide more comprehensive coverage for future *citation function*-related analysis tasks.

2.2. Argumentative Structure of The Research Paper

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 thesis since the proposed annotation scheme in this thesis naturally contains argumentative labels.

Argumentative structures can be applied to a section-level or sentence-level category. (Sollaci & Pereira, 2004) used 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. (Simone 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 (Simone; 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 eight labels built based on 40 Computer Graphics papers.

2.3. A new Annotation Scheme Development

The proposed annotation scheme for *citation functions* in this thesis is developed by following several steps. First, the top-down and bottom-up analyses are required. 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 thesis uses a dataset from well-parsed sentences from arXiv (Färber et al., 2018). This study 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 2.2.

Table 2.2. The proposed annotation scheme for citation functions in this thesis.

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:	however, this coherence metric is widely used for the cs scenario due to its simplicity <citation>.

	explaining the significance of the research topic, theory, principle, concept, topic, problem, etc.	
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 thesis, 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>.
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>.

2.4. Citation Scheme Comparison

As part of scheme development, a label comparison is performed between the scheme proposed in this thesis and existing schemes. As mentioned before, the existing schemes consist of *citation functions* and argumentative structures. Through comparison, this study shows the compatibility and contribution of the proposed scheme. In Table 2.3, N/A marks indicate the newly proposed labels of the scheme that were not accommodated in existing works. The comparison reveals that the 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 (Simone Teufel et al., 1999) and (Simone; 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.

2.5. Annotation Strategy

Annotation experiments are the last part of scheme development. Two CS master's degree graduates (annotators) are employed 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 instances to be labeled. The *Inter-annotator Agreement* and Kappa values (Cohen, 1960) are used 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 (J. Wang et al., 2019).

Table 2.3. Comparison between the proposed scheme in this thesis and existing schemes.

Fine-Grained Classes of the proposed Scheme in this thesis	**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 & Schäfer, (2011): Background; Jurgens et al., (2018): Background; Zhao et al., (2019): Introduce; Cohan et al., (2019): Background; Pride & 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 & Knoth, (2020): motivation; Teufel et al., (2006): Pbas; Dong & 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;
citing_paper_use	Pride & Knoth, (2020): use; Tuarob et al., (2019): use, extend; Teufel et al., (2006): Puse; Dong & 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 & 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 & Knoth, (2020): future	N/A	Teufel et al., (2009): FUT; Fisas et al., (2015): Future Work	N/A
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 & Knoth, (2020): similarities; Teufel et al., (2006): Psim; Casey et al., (2019): A-CW-SIM	Dong & Schäfer, (2011): Comparison, Valenzuela et al., (2015): Comparison Zhao et al., (2019): compare, Cohan et al., (2019): Result Comparison Su et al., (2019): Compare and Contrast	N/A	Teufel et al., (1999): contrast; Teufel et al., (2009): CODI
contrast	Pride & Knoth, (2020): contrast; Teufel et al., (2006): CoCoGM, CoCoR0; Hernández-Álvarez et al., (2016): contrast; Bakhti et al., (2018): contrast		N/A	
other_multiple_intent	N/A	N/A	N/A	N/A
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

2.6. Annotation Experiment Results

The annotation experiment results contain raw agreement and Kappa values. The confusion matrix in Table 2.4 and Table 2.5 shows raw agreements 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 the proposed scheme, the obtained Kappa values are competitive compared with previous works, e.g., (A. Casey et al., 2019) with 0.77, (Simone Teufel et al., 2006) with 0.72, (Dong & Schäfer, 2011) with 0.757, and (Zhao et al., 2019) with 0.47.

Table 2.4. Confusion matrix for Inter-annotator Agreement on five *coarse* labels.

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**

Table 2.5. Confusion matrix for Inter-annotator Agreement on fine-grained labels.

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
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
other_other	0		2	1	2	0	0	0	3	0	0	0	1	0	0	2	0	1	0	1	0	**1**

This study highlights 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 annotator II labeled as *citing_paper_corroboration*. In this case, both labels are part of the *coarse* label *citing paper work* and the 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.

2.7. Chapter Summary

This chapter has developed a new labeling scheme of *citation functions* based on multi-fields CS research papers. The motivation behind this development is to address the limitation of existing schemes which contained few numbers of labels and were developed based on limited research scopes. The proposed scheme in this thesis is developed by following two steps, i.e., 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. After conducting these steps, the final scheme consists of 5 *coarses* 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.

Chapter 3 – The Development of a New Dataset of Citation Functions

This chapter explains the method to develop a new dataset of *citation functions*. The motivation behind this development is to address several limitations in existing *citation functions* datasets which contain few instances, limited number of labels, and the labels were built using narrow research fields. The solution proposed in this thesis is a semi-automatic approach based on two types of datasets. The first type contains 5,668 manually labeled instances to develop a new labeling scheme of *citation functions*, and the second type is the final dataset that is built automatically. The proposed dataset is developed based on a new labeling scheme of *citation functions* has been develop in this thesis, consisting 5 *coarses* labels and 21 *fine-grained* labels. To realize this, there are several text classification scenarios to be implemented, such as classical Machine Learning (ML), Deep Learning, and Word Embedding. In addition, this thesis proposed to implement Active Learning (AL) as a low resource scenario.

3.1. Existing Dataset of Citation Functions

The existing works which performed *citation functions* 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 (X. Li et al., 2013) to propose *coarse* annotation with sentiment labels. Because of the limitation of labeled instances, (Dong & 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-Alvarez et al., 2017). 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 functions* and provenance classification. This task was implemented using multitask learning. Sharing a similar multitask setting, while (Bakhti et al., 2018) also used 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 & Knoth, 2017) used influential and non-influential citations to find

highly predictive features. The classification in this work was performed using RF. Next, (M. 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., 1999) with re-annotation and developed a model using SVM. Following this, (Roman et al., 2021) used the citation context dataset from CORE[10]. This research applied BERT, depending on the three labels proposed by SciCite (Cohan et al., 2019).

Table 3.1. Existing datasets of citation functions, their source papers, and the number of citing sentences.

No.	Research Paper	Sample Papers	Number of Labeled Instances
1	(Simone Teufel et al., 2006)	116	2,829
2	(Dong & Schäfer, 2011)	122	1,768
3	(X. Li et al., 2013)	91	6,355
4	(Valenzuela et al., 2015)	20,527	465
5	(Hernández-Álvarez et al., 2016)	85	2,092
6	(Jurgens et al., 2018)	185	1,969
7	(Bakhti et al., 2018)	?	8,700
8	(A. Casey et al., 2019)	95 related work sections	1,806
9	(Rachman et al., 2019)	Dataset 1: 160 Dataset 2: 50	Dataset 1: 2,475, Dataset 2: 1,153
10	(Su et al., 2019)	n/a	1,432
11	(Zhao et al., 2019)	39,601	3,088
12	(Cohan et al., 2019)	6,627	11,020
13	(Tuarob et al., 2019)	8,063	8,796
14	(David Pride & Knoth, 2020)	883	11,233
15	(Roman et al., 2021)	?	10 million

[10] https://core.ac.uk/

Table 3.1 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 were 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.

3.2. Building Citation Functions Dataset through Classification

This section describes how the proposed dataset is developed using a semiautomatic approach. The entire system consists of three stages. The first stage is annotation scheme development which was created in the previous part of this thesis. In this stage, the existing labels of *citation functions* are identified and reviewed. 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 thesis uses several classification scenarios to build these models. The first scenario is implemented using a classical deep learning method. Next, this thesis applies 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 3.1 depicts how the proposed dataset is developed.

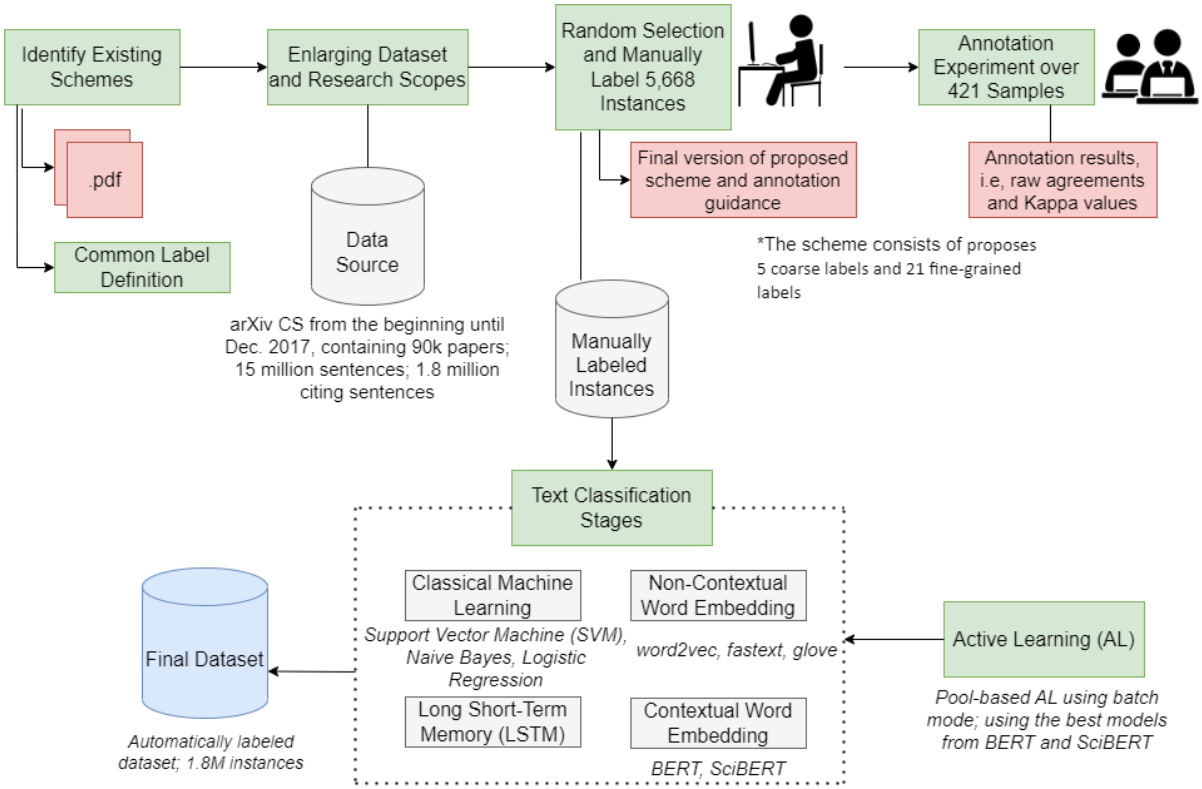


Figure 3.1. Development of the semiautomatic dataset of citation functions.

3.2.1. Text Classification Strategy

The 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 5,668 instances. Here, this thesis evaluates 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 the deep learning method. Considering the few numbers of labeled instances, it is worth applying pretrained word embeddings. This thesis implements 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 $1e^{-5}$ to $5e^{-5}$, 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, this thesis uses *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. 3.2 depicts the distribution of the development dataset for all classification strategies.

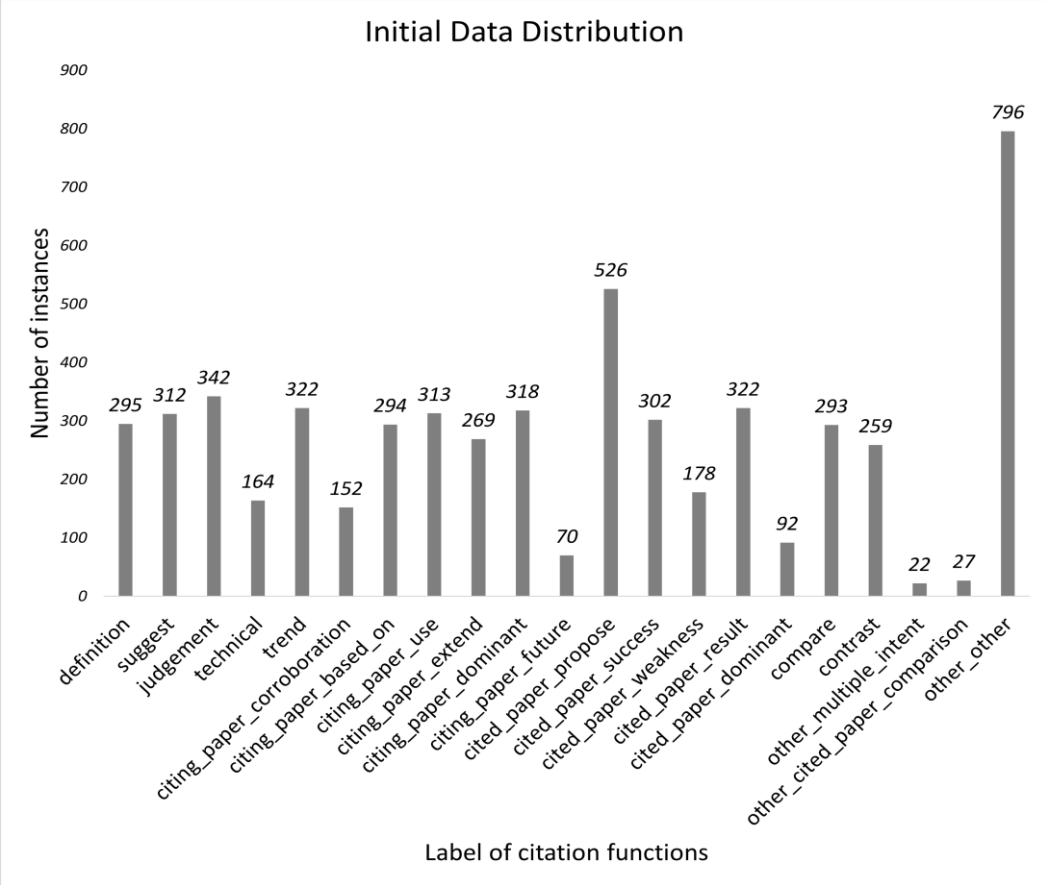


Figure 3.2. Development (initial) labeled instance distribution.

3.2.2. Active Learning Strategy

Active learning (AL) is subfield of Machine Learning which allows the algorithm to choose 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 3.3 shows the pool-based scenario as the most common scenario of the AL as presented by (Settles, 2010). (Lewis & 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.

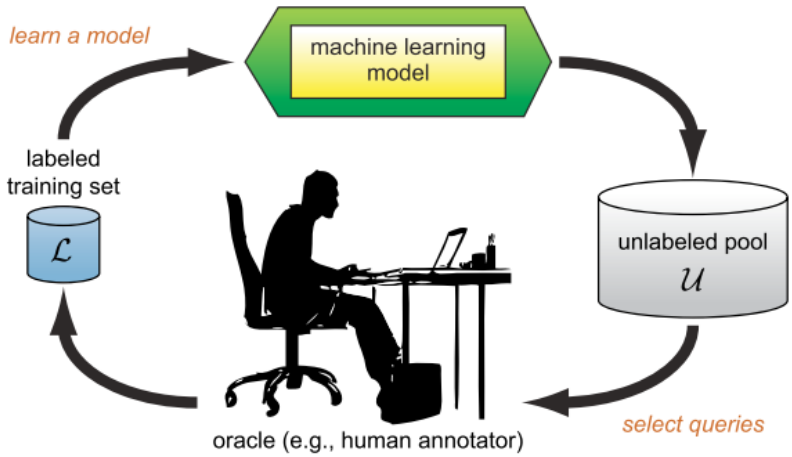


Figure 3.3. Pool-based active learning scenario.

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 & 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.

$$x_{LC}^* = \underset{x}{\operatorname{argmax}}\, 1 - P_\theta(\hat{y}|x) \qquad \text{Equation 3.1}$$

- *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):

$$x_M^* = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x) \qquad \text{Equation 3.2}$$

- *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.

$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \qquad \text{Equation 3.3}$$

AL has been successfully used to reduce the manual labeling effort. This thesis implements the pool-based AL strategy using a batch mode as illustrated in Figure 3.4. 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 1,000 new instances from *U* that are gradually added to *L*. For batch selection, this thesis compares three sampling approaches, i.e., *least confident*, *max-margin*, and *entropy*. Note that this thesis follows the AL strategy proposed by (Ein-Dor et al., 2020; Hu et al., 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.

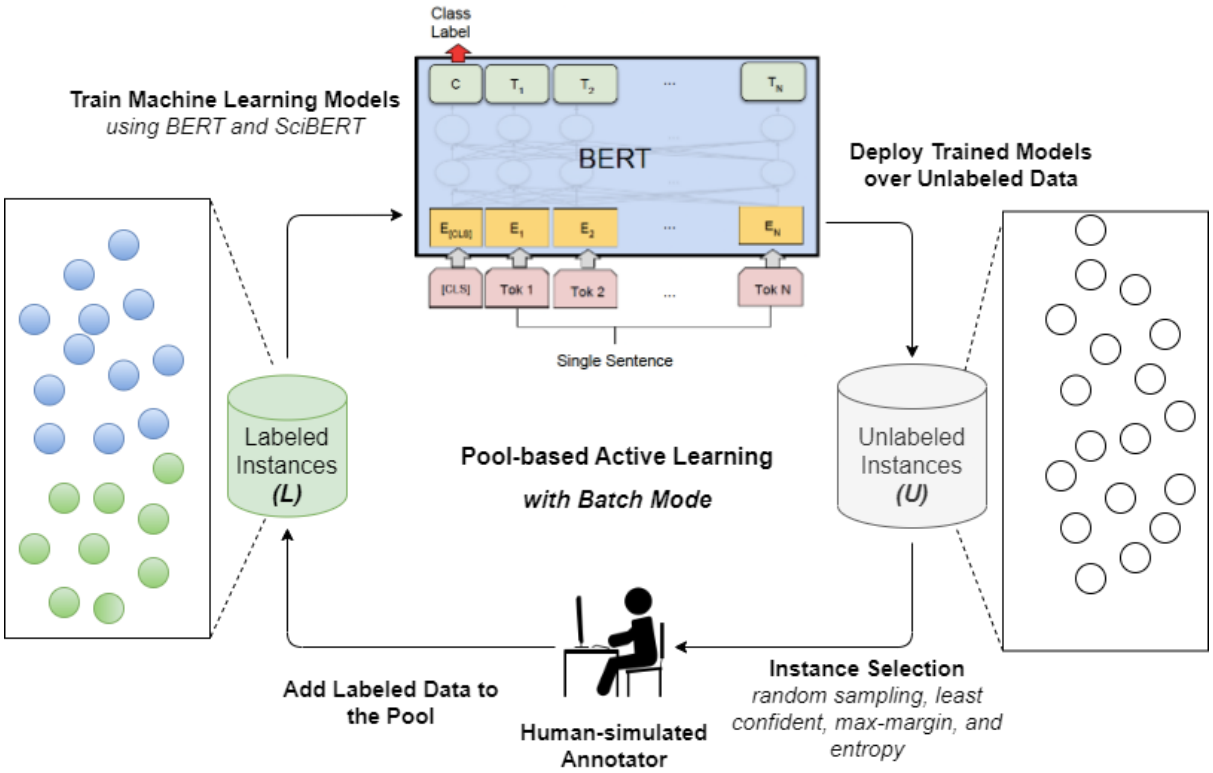


Figure 3.4. Pool-based active learning used in this thesis.

3.3. Experiment Results of Classification Scenarios

This section demonstrates the results of the classification experiments. The results are organized based on the classification stages.

3.3.1. Filtering Stage Results

Table 3.3 shows 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 only 75.99%. Notably, classical classifiers achieved almost similar accuracies of 85%. But if looking 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 3.4 depicts the parameters used in the *filtering* stage.

Table 3.2. The best testing results of each classification technique for the *filtering* stage.

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**

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

Table 3.3. 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	$2e^{-5}$; batch 64; imbalance
SciBERT	$3e^{-5}$; batch 32; balance

Looking at the performance of each label in Figure. 3.5, all performance metrics in the *noother* label are lower than *other* label. There are extreme cases where the *noother* label has zero values as in Naïve Bayes, word2vec, and fasttext. Two methods, BERT and SciBERT, relatively have balanced proportions compared with other methods.

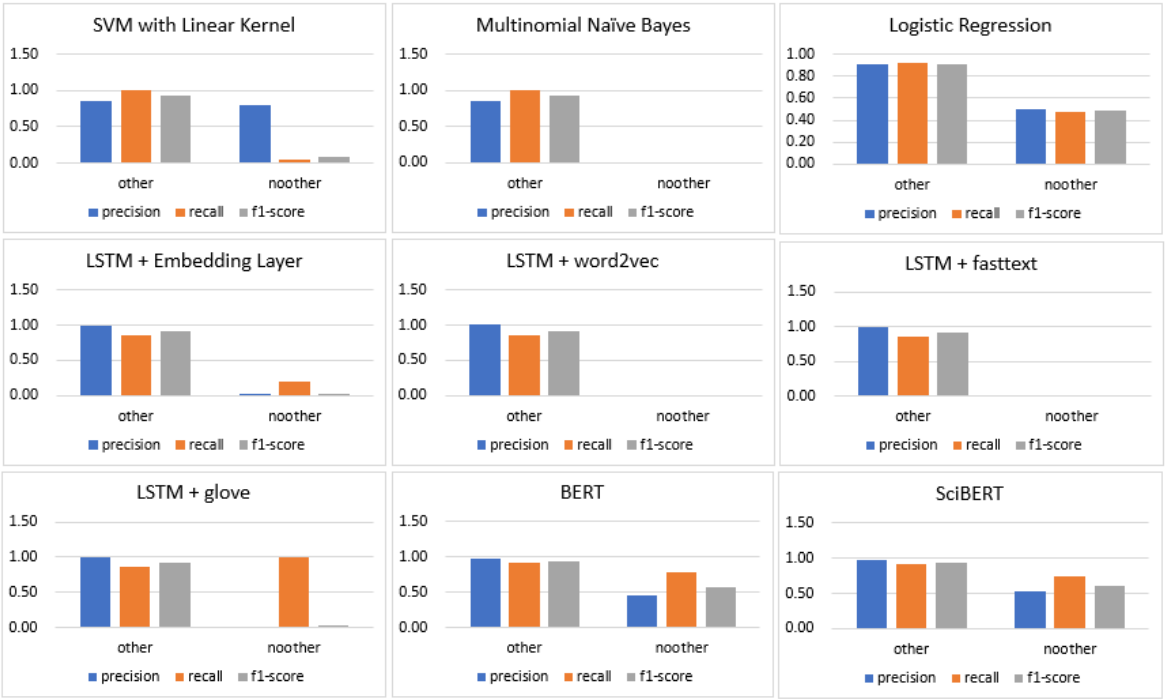


Figure 3.5. The performance metrics of individual class in the *filtering* stage.

**The x-axis depicts the classes and their performance metrics, and the y-axis depicts the performance values.*

3.3.2. Fine-grained Results

As predicted, the performance in this stage will be lower than that in the *filtering* stage. Table 3.5 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 looking at the individual performance, four approaches i.e., embedding layer, word2vec, fasttext, and glove show poor results (Figure. 3.6). 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 3.6. The full performance comparison of BERT and SciBERT in the *filtering* and *fine-grained* stages is shown in Figure. 3.7.

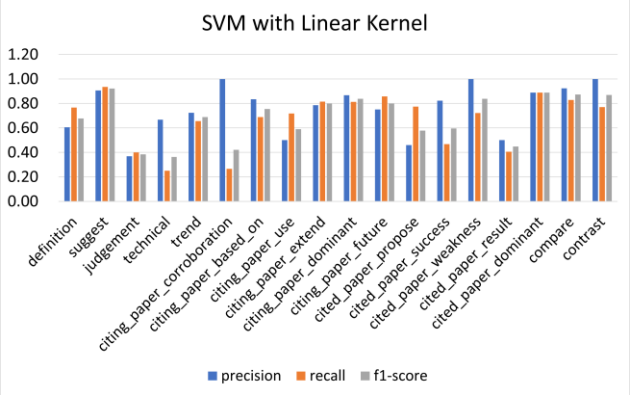
Table 3.4. The best testing results of each classification technique for *fine-grained* labels.

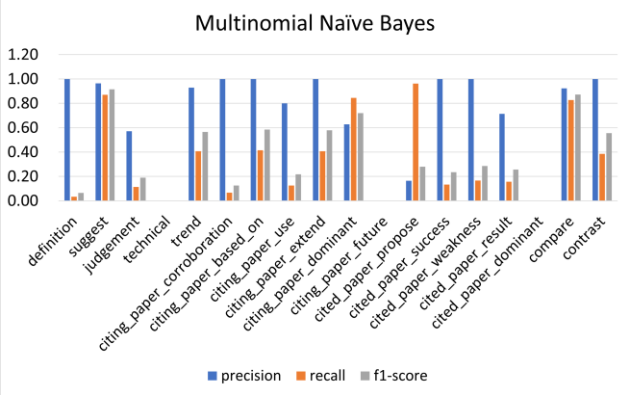
Methods	**Accuracy**	**Macro avg precision**	**Macro avg recall**	**Macro avg f1**
SVM	67.29	75.57	66.79	68.49
Naïve Bayes	57.55	74.06	50.94	52.87
Logistic Regression	69.98	71.87	70.23	70.53
LSTM + Embedding Layer	13.87	10.22	8.09	6.02
LSTM + word2vec	10.97	7.73	2.29	3.45
LSTM + fasttext	14.49	10.02	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 (%).*

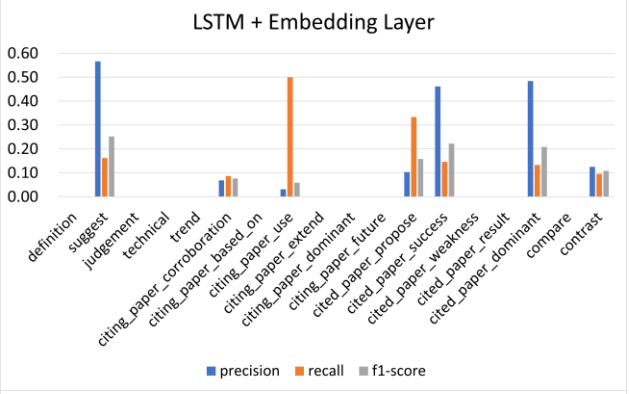
Table 3.5. Hyper-parameters used in the *fine-grained* labels.

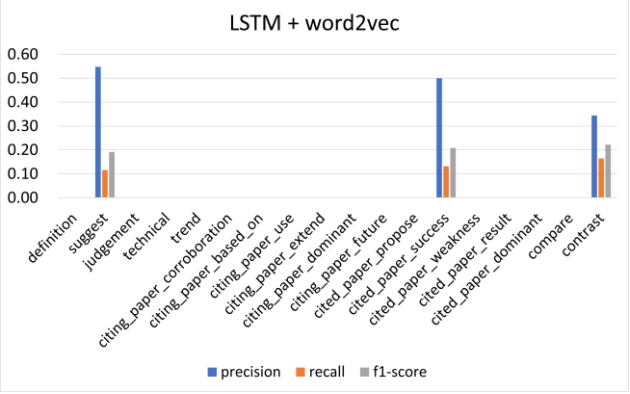
Techniques	**Parameters**
SVM	ngram_range: (1, 2); TF/IDF; imbalance; kernel=linear;
Naïve Bayes	ngram_range: (1, 2); bag of word; imbalance;
Logistic Regression	C: 1; penalty: l1; ngram_range: (1, 2); TF/IDF; imbalance; solver='liblinear'
LSTM + Embedding Layer	epoch 3; batch 32; imbalance; optimizer=adam; 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	$3e^{-5}$; batch 32; imbalance
SciBERT	$3e^{-5}$; batch 32; balance

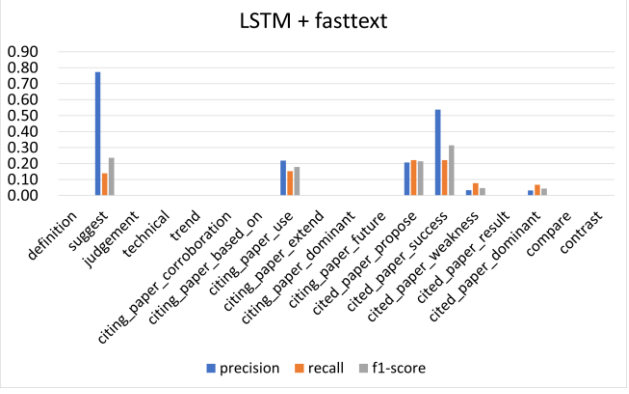

SVM with Linear Kernel
1.20
1.00
0.80
0.60
0.40
0.20
0.00
definition
suggest
judgement
technical
trend
citing_paper_corroboration
citing_paper_based_on
citing_paper_use
citing_paper_extend
citing_paper_dominant
citing_paper_future
cited_paper_propose
cited_paper_success
cited_paper_weakness
cited_paper_result
cited_paper_dominant
compare
contrast
precision
recall
f1-score


Multinomial Naïve Bayes
1.20
1.00
0.80
0.60
0.40
0.20
0.00
definition
suggest
judgement
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
precision
recall
f1-score


Logistic Regression
1.20
1.00
0.80
0.60
0.40
0.20
0.00
definition
suggest
judgement
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
precision
recall
f1-score


LSTM + Embedding Layer
0.60
0.50
0.40
0.30
0.20
0.10
0.00
definition
suggest
judgement
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
precision
recall
f1-score


LSTM + word2vec
0.60
0.50
0.40
0.30
0.20
0.10
0.00
definition
suggest
judgement
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
precision
recall
f1-score


LSTM + fasttext
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
definition
suggest
judgement
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
precision
recall
f1-score

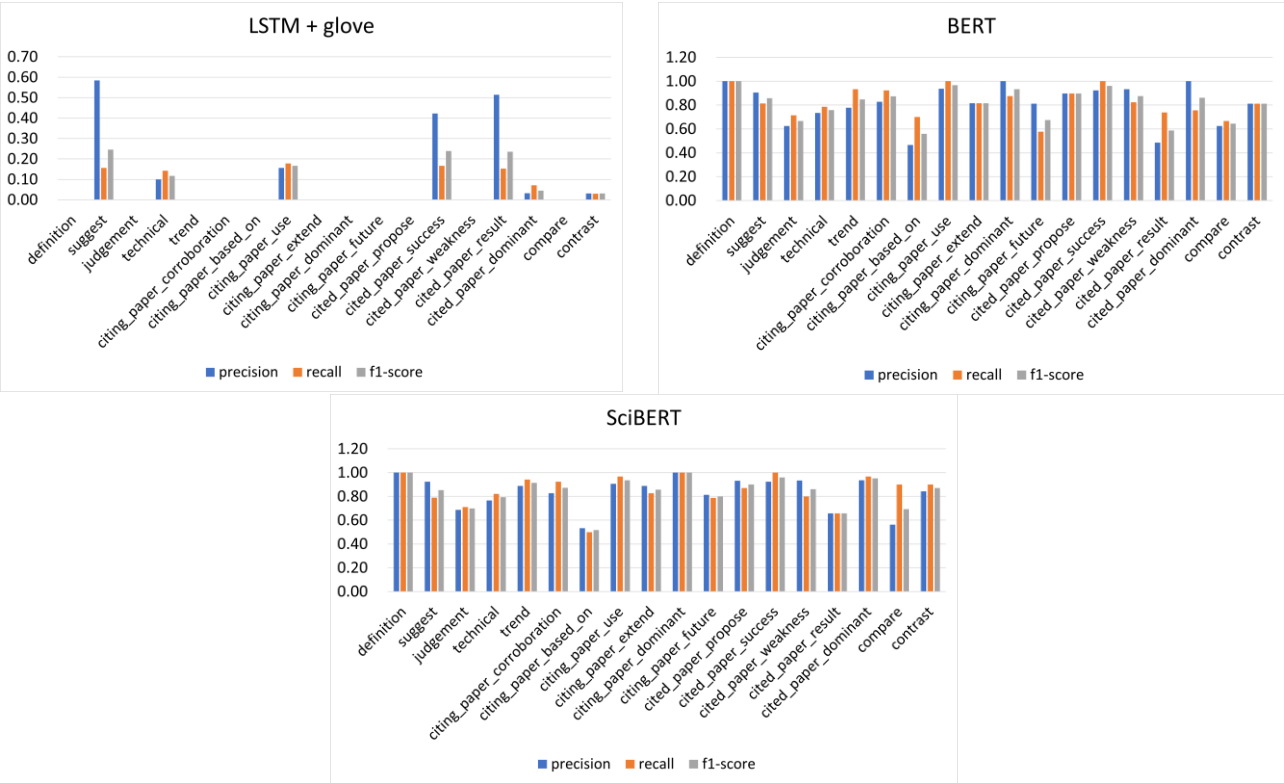


Figure 3.6. Performance metrics of each class in the fine-grained stage.

**The x-axis depicts the classes and their performance metrics, and the y-axis depicts the performance values.*

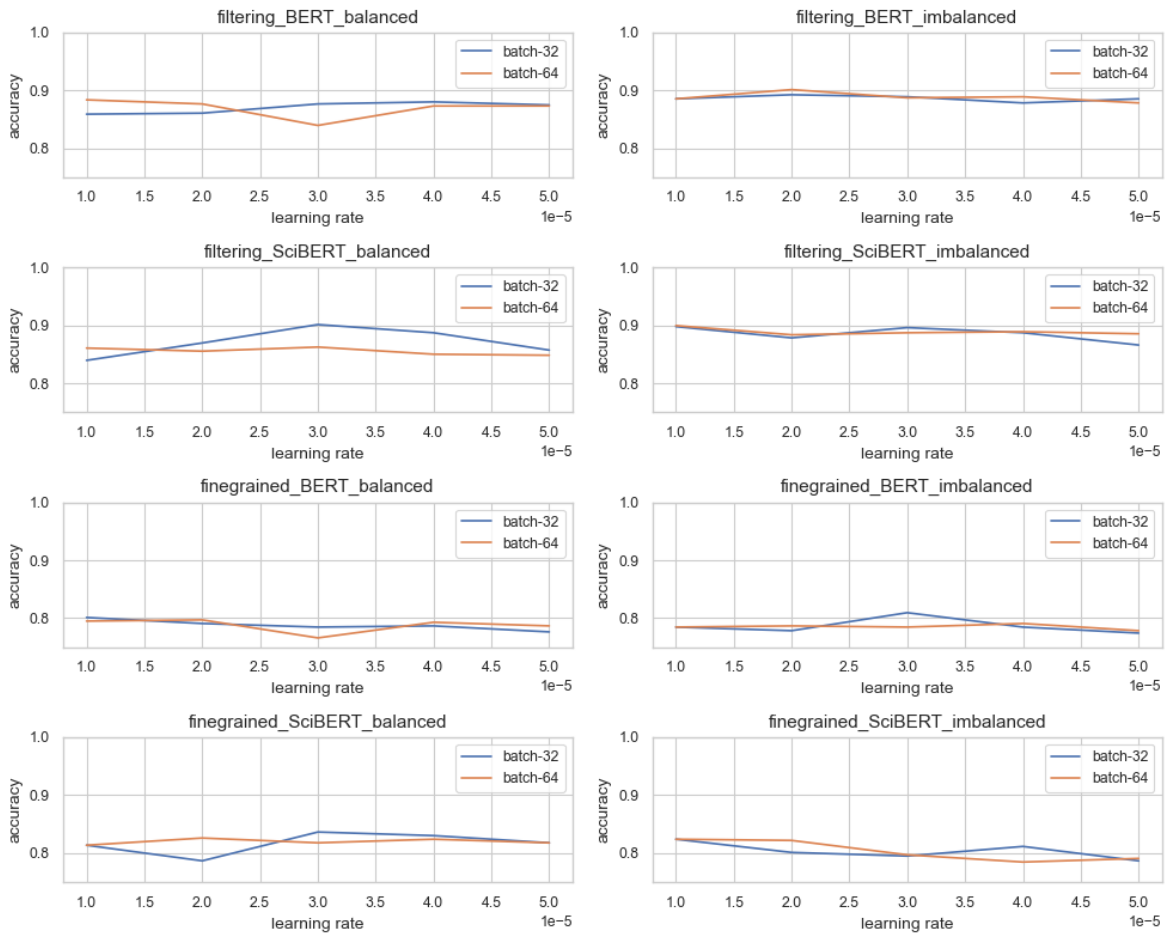


Figure 3.7. BERT and SciBERT performance comparison in filtering and fine-grained stages depend on learning rates and batches.

3.3.3. 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 $2e^{-5}$, batch size of 64, and imbalanced distribution in the BERT-based AL. For the SciBERT experiments, the best parameters were learning rate of $3e^{-5}$, batch size of 32, and balance distribution. BERT-based

fine-grained experiment implemented the AL strategies based on learning rate of $3e^{-5}$, batch size of 32, and imbalanced distribution. For SciBERT, the parameters used were learning rate of $3e^{-5}$, batch size of 32, and balanced distribution.

STATISTICALLY SIGNIFICANT TEST

Since this thesis implements two classification scenarios, i.e., non-AL and AL, this thesis 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 2x2 contingency table.

Table 3.6. The 2x2 Contingency Table of the McNemar's test.

	classifier 2 - correct	classifier 2 - wrong
classifier 1 - correct	a	b
classifier 1 - wrong	c	d

According to Table 3.6, the test statistic is calculated as follows:

$$X^2 = \frac{(b-c)^2}{(b+c)} \qquad \text{Equation 3.4}$$

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, this thesis needs 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, this thesis can *Reject Null Hypothesis* because the two compared classifiers are significantly different. The *p-value* is calculated as follows:

$$p - value = 1 - cdf(X^2) \qquad \text{Equation 3.5}$$

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

FILTERING STAGE

AL-based performance in the *filtering* stage is depicted in Table 3.6. 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 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 Figure. 3.7.

Table 3.7. The best result in the *filtering* stage for AL strategies.

Classification Strategies	**max_margin**		**entropy**		**least_confident**		**random_sampling**	
	Queries	**Accuracy (%)**	**Queries**	**Accuracy (%)**	**Queries**	**Accuracy (%)**	**Queries**	**Accuracy (%)**
BERT	450	88.71	500	88.88	1,000	**90.29**	900	88.35
SciBERT	900	88.00	850	89.59	800	88.88	650	89.41

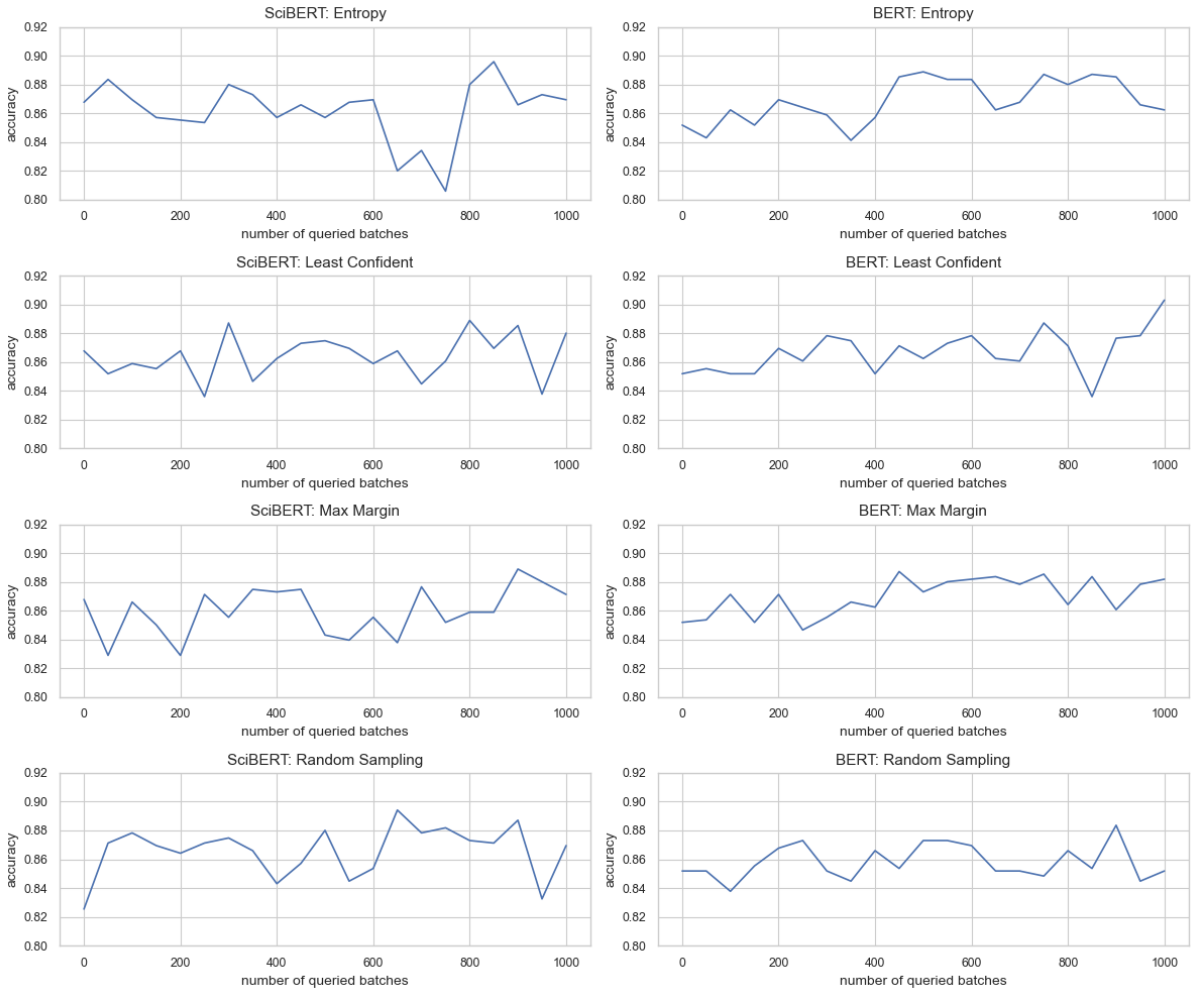


Figure 3.8. Result comparison of AL strategies on the *filtering* stage using BERT and SciBERT with four sampling approaches. The data splitting scenario is 1,039 (testing), 4,534 (simulating *L* and *U*), and 453 (seed).

As the AL-based strategy in the *filtering* stage achieved slightly higher accuracy (90.29%) compared to the non-AL strategy (90.12%), a statistically significant test based on the McNemar approach was conducted. 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, the experiment measured alternative metrics as shown in Table 3.8 as performed in the non-AL setting. Even failed to reject the null hypothesis, the results 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 3.8. 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

FINE-GRAINED STAGE

The AL-based performance in *fine-grained* classification is depicted in Table 3.8. 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 1,000 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 Figure. 3.9.

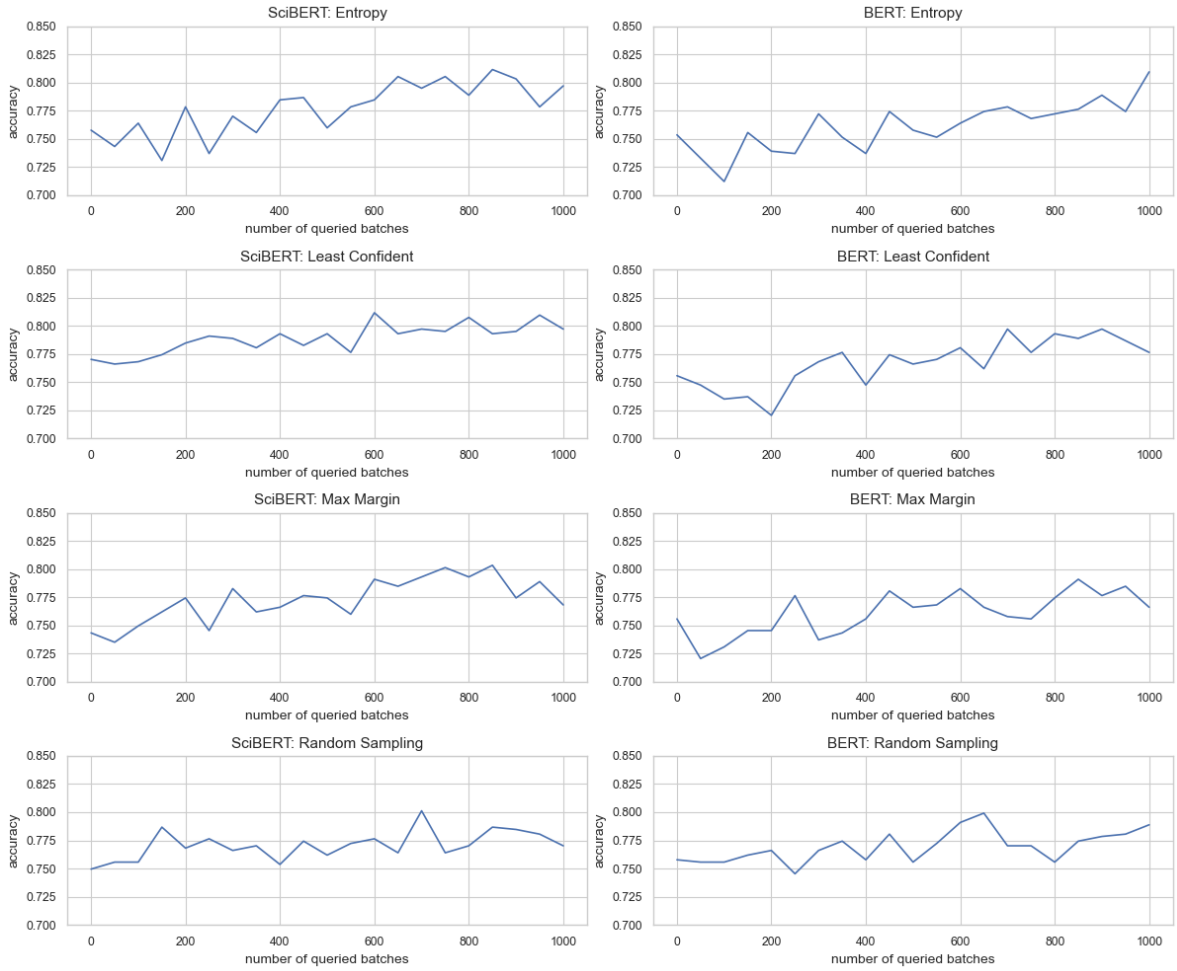


Figure 3.9. Result comparison of AL strategies for *fine-grained* classification using BERT and SciBERT with four sampling approaches.

** The data splitting scenario is 1,039 (testing), 3,858 (simulating L and U), and 771 (seed).*

Table 3.9. The best result of *fine-grained* AL strategies.

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

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 3.10.

Table 3.10. Detailed performance metrics of the best accuracy in the AL strategy.

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

**All metrics are measured by percentage (%).*

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.

Table 3.11. The distribution of new dataset of citation function.

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**

Finally, this study uses the best models to classify unlabeled *citing sentences*. Table 3.11 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 5,439. 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 Figure. 3.2) 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.

3.4. Chapter Summary

This chapter developed a dataset of *citation functions* consisting of 1,840,815 labeled instances. The dataset was built using a semiautomatic approach. Specifically, the proposed method trained machine learning models on manually labeled data and use these models to label unlabeled instances. The proposed scheme was developed through top-down analysis, bottom-up analysis, and annotation experiments. Besides the achieved 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 thesis 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, it is important 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 the 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.

Chapter 4 – Paper Quality Prediction

This chapter explains the prediction method of paper quality that has been proposed in this thesis. The prediction method is intended to support Technology Assisted Peer Review (TAPR) to reduce the review burden. This thesis offers a prediction method relying only on the paper itself to address applicability and fairness issues that exist in most existing works. The applicability issues happen when the purpose of prediction method is to reduce the review burden, but the method is still involving the reviewers' comments as prediction features. In addition, the fairness issue exists because the developed predicting method directly estimated the final review decision (accepted or rejected) even though in fact the final decision is made by Editor based on multiple factors. Handling these issues, this thesis developed a prediction method covering three tasks, two are classification tasks and the last one is a regression task. The classification tasks focus on predicting the final review decision (*accepted-rejected*) and estimating the paper quality (*good-poor*), and a regression task to predict the review scores. By predicting whether the submitted paper is good or poor together with its review scores, objectively assessing the paper quality is closely estimated.

4.1. Existing Works on Paper Quality Prediction

This section presents existing works on two research focuses as follows: (a) classification tasks, comprising the final review decision and the paper quality prediction, and (b) regression tasks for predicting review scores. These works are reviewed by focusing on three aspects, i.e., the prediction tasks, the datasets, and the prediction features. Finally, this thesis highlights the limitations of how existing prediction methods were developed and illustrates the contribution of this thesis to this research area.

The ICLR is the most widely adopted source for discussing the dataset used to make predictions. This trend is because the ICLR provides both accepted and rejected papers accompanied with peer-review information, such as review comments and review scores. In this study area, the dataset published by (Kang et al., 2018) is the most cited work, which compiled numerous peer-review datasets comprising ICLR, arXiv, Association for Computational Linguistics (ACL), and Conference on Computational Natural Language Learning (CoNLL). However, only two works used the non-ICLR dataset, such as (A. J. Casey et al., 2019) which used the 94 Related Work section of the ACL dataset, and (Ribeiro et al., 2021) which used paper collections obtained from the Artificial Intelligence (AI) Conference (2013 and 2019) and Robotics (2015 and 2019).

Two major categories of classification features are used in the existing works in the classification tasks. The first category is classifying features developed based on the manuscript's content. In this category, the proposed features range from lexical features to word representation methods. Alternatively, the second category is classifying the features by employing review comments (most existing works fall into this category). Additionally, most existing works treated the prediction as a binary *accepted-rejected* classification task. For example, studies proposed more than two classes, as in (K. Wang & Wan, 2018) which used two and three labels for *accepted-rejected* and accepted-borderline-rejected, respectively, and in (A. J. Casey et al., 2019) with three classes of good-average-poor. Conversely, most existing studies predicted the aspect review scores in the regression task as the structured summary reflecting the manuscripts' strengths and weaknesses. Therefore, this aspect of the review scores can contain several points, e.g., impact, recommendation, substance, clarity, etc., as stated in (Kang et al., 2018). Additionally, two existing studies proposed the final review scores as in (Ghosal et al., 2019; Ribeiro et al., 2021).

The literature review poses some limitations in most existing publications. First, the crucial role of *citation functions* was omitted from being addressed in assessing the paper's quality. Second, existing studies did not provide what the manuscript's aspects or sections are important to predict its quality. Third, the unfairness of using review comments as prediction features and using only accuracy as the only metric biased toward the majority class. Fourth, the bias of predicting only *accepted-rejected* due to the final review decision relies on multiple factors. Therefore, this thesis develops a prediction method that depends only on the manuscript's content, particularly using the *citation functions* obtained from *citing sentences* to resolve these challenges. This study proposes creating two additional prediction features, *regular sentences* and *reference-based* features. The paper majorly aims to predict the paper quality (*good-poor*) and the review scores. The final review decision is covered as well for comparison purposes. Accordingly, the prediction features address the limitation of determining the most influential part of the manuscript to predict its quality using several ML and FS methods.

Interestingly, the study by (K. Wang & Wan, 2018) conducted experiments on the three classes of accepted, borderline, and rejected, and the two classes accepted and rejected by eliminating the borderline papers. Although eliminating the borderline papers improved the prediction performance, this becomes inapplicable in the entire peer-review process. Additionally, when a reviewer judges a paper as borderline, it does not mean that the other two reviewers judge it as the same since the submitted manuscripts are reviewed by three reviewers and have three different review scores. Due to this reason, this study prefers to use the average review scores to determine whether a paper is good or poor (further explanation of this issue is presented in the subsequent section). Casey et al. (A. J. Casey et al., 2019) proposed good, average, and poor as final quality decisions in which the labels are determined by the annotator and not by conference reviewers or editors in a study with the same three-class boundaries. Tables 4.1 and 4.2 show the details of the existing studies.

Table 4.1. Existing studies on final review decision and paper quality prediction.

Paper	Title	Dataset for Prediction	Feature	Focused Task
(Kang et al., 2018)	A Dataset of Peer Reviews (PeerRead): Collection, Insights, and NLP Applications	ICLR 2017, ArXiv	hand-engineered coarse and lexical features.	Accepted-rejected
(K. Wang & Wan, 2018)	Sentiment analysis of peer-review texts for scholarly papers	ICLR 2017–2018	review comments.	Accepted-rejected (2 classes); and accepted-borderline-rejected (3 classes)
(Jen & Chen, 2018)	Predicting Conference Paper Acceptance	Using ICLR 2017	pre-defined features.	Accepted-rejected
(A. J. Casey et al., 2019)	Can Models of Author Intention Support Quality Assessment of Content?	94 Related Work section from ACL papers.	10 author's intention in the related work section	Good-average-poor (3 classes)
(Ghosal et al., 2019)	Deep Sentipeer: Harnessing sentiment in review texts to recommend peer-review decisions	ICLR 2017, ACL 2017, CoNLL 2016	the paper, review comment, review sentiment.	Accepted-rejected
(Ghosh et al., 2020)	Conference Paper Acceptance Prediction (Acceptometer)	ICLR 2017	the paper, review comment.	Accepted-rejected
(Skorikov & Momen, 2020)	Machine learning approach to predicting the acceptance of academic papers	NIPS, ICLR, ACL, CoNLL, ArXiv (treated as single dataset)	pre-defined features.	Accepted-rejected
(Maillette de Buy Wenniger et al., 2020)	Structure-tags improve text classification for scholarly document quality prediction	ArXiv	tag structure of the paper.	Accepted-rejected
(Ciloglu & Merdan, 2020)	Big Peer-Review Challenge	NIPS, ICLR, ACL, CoNLL, ArXiv (treated as single dataset)	the paper, review comment.	accepted-rejected
(Joshi et al., 2021)	Conference Paper Acceptance Prediction: Using Machine Learning	ICLR 2017	pre-defined coarse and lexical features.	Accepted-rejected
(Vincent-Lamarre & Larivière, 2021)	Textual analysis of artificial intelligence manuscripts reveals features associated with peer-review outcome	NIPS, ICLR, ACL, CoNLL, ArXiv (treated as single dataset)	bag of words.	Accepted-rejected
(Bao et al., 2021)	Predicting Paper Acceptance via Interpretable Decision Sets	ICLR 2017, ArXiv	pre-defined features.	Accepted-rejected
(Ribeiro et al., 2021)	Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis	AI Conference 2013 & 2019, and Robotics 2015 and 2019	review comment.	Accepted-rejected
(Bharti et al., 2021)	PEERAssist: Leveraging on Paper-Review Interactions to Predict Peer-Review Decisions	ICLR 2017–2020	the paper, review comment, review sentiment.	Accepted-rejected

(Fytas et al., 2021)	What Makes a Scientific Paper be Accepted for Publication?	ICLR 2017	review comment, meta review.	Accepted-rejected
(Pradhan et al., 2021)	A deep neural architecture based meta-review generation and final decision prediction of a scholarly article	ICLR 2017–2019	review comment.	Accepted-rejected

Table 4.2. Existing works on review score prediction.

Authors	Paper Title	Category	Dataset	Features
(Kang et al., 2018)	A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications	Aspect review score prediction	ACL 2017; ICLR 2017	Review comments
(Ghosal et al., 2019)	DeepSentiPeer: Harnessing Sentiment in Review Texts To Recommend Peer-Review Decisions Tirthankar	Final review score prediction	ICLR 2017; ACL 2017; CoNLL 2016	Review comments
(Stappen et al., 2020)	Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus	Aspect review score prediction	Interspeech 2019	Review comments
(Q. Wang et al., 2020)	ReviewRobot: Explainable Paper-Review Generation based on Knowledge Synthesis	Aspect review score prediction	ACL 2017	Review comments
(J. Li et al., 2020)	Multi-task Peer-Review Score Prediction	Aspect review score prediction	ICLR 2017 and ACL 2017	Paper text
(Ribeiro et al., 2021)	Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis	Final review score prediction	AI Conference 2013 & 2019; Robotics 2015 & 2019	Review comments

A list of abbreviations: Conference on Neural Information Processing Systems (NIPS), Association for Computational Linguistics (ACL), Computational Natural Language Learning (CoNLL), Artificial Intelligence (AI)

4.2. The Prediction Method

This section briefly describes the stages used to build the prediction method proposed in this thesis, as shown in Figure 4.1. The prediction method follows several stages: The first stage discusses the research papers' data source, which is a paper acceptance dataset. The second stage explains three predictors having classification and regression features due to the system being treated as classification and regression problems. These predictors are *citing sentence*

predictors developed based on the labeling scheme of *citation functions*, *regular sentence* predictors created by applying the label of *citation functions* to non-citation text, and *reference-based* features constructed by identifying the source of citations. Finally, the final stage explains the proposed prediction scenarios and evaluations.

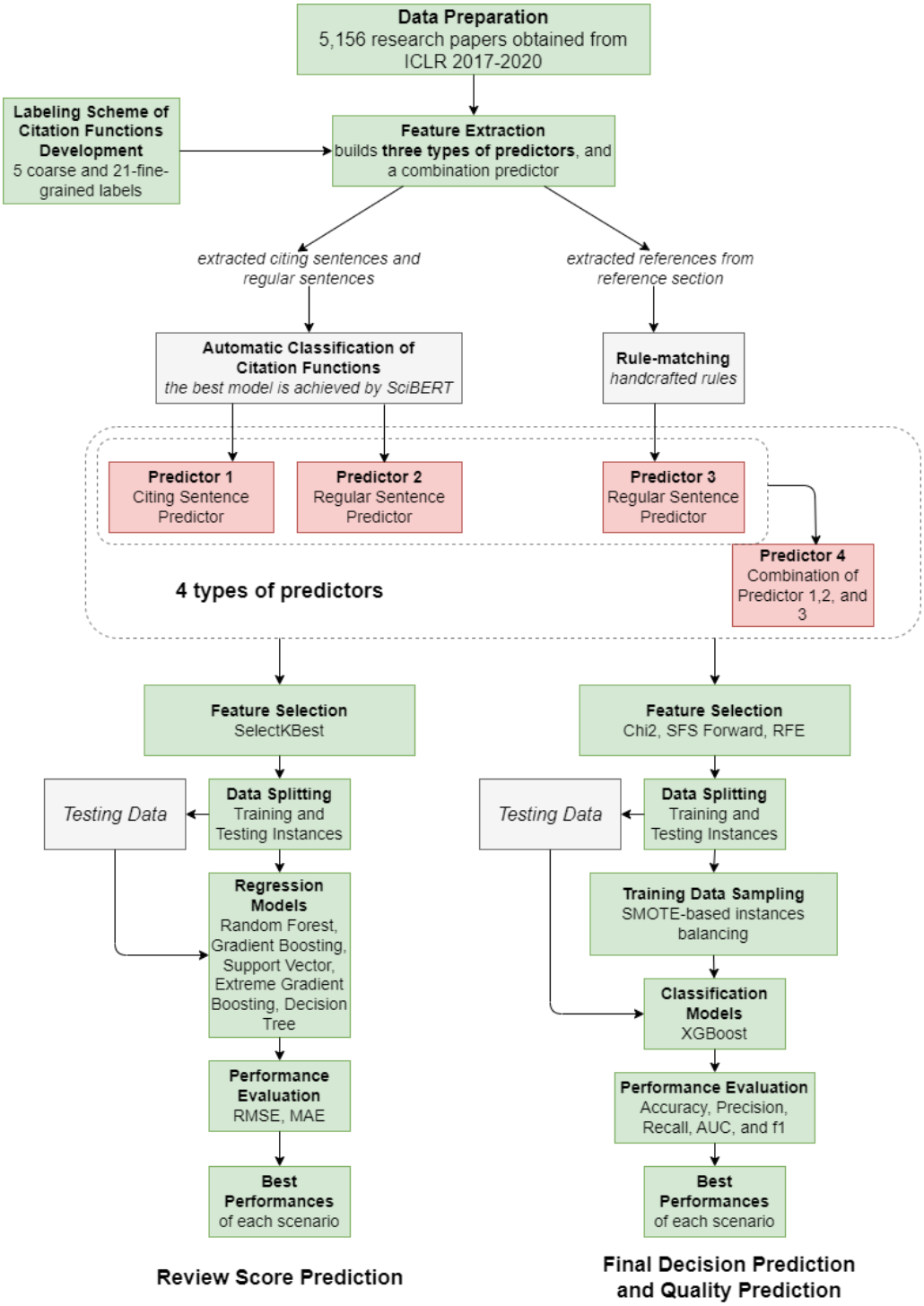


Figure 4.1. The general architecture of the proposed method for both classification tasks and regression task.

4.1.1. Citing Sentence Predictor

The *citing sentence* predictor is the first proposed and main technique to estimate all prediction tasks. This predictor is developed based on the *citation functions*, which explains why the author of the research papers cited previous works. Therefore, this thesis uses the labeling scheme of *citation functions* developed in the previous stage (Basuki & Tsuchiya, 2022) comprising 5 *coarse* and 21 *fine-grained* labels. The scheme of *citation functions* was developed using a research paper dataset from (Färber et al., 2018), containing 90,278 parsed papers from arXiv Computer Science (CS) from January 1993 to December 31, 2017. Furthermore, the *coarse* labels are defined for representing the general idea of the *citation functions* and *fine-grained* labels to develop a detailed version of the labels. Moreover, all these labels are applied as features, and one more feature is included to represent the number of *citing sentences* in each paper. The features are developed by classifying all *citing sentences* in the ICLR dataset using ML and calculating the labels contained in each paper. Finally, the features are encoded as *c0 to c19*, as shown in Table 4.3.

Table 4.3. The coarse and fine-grained labels of citation functions as a list of features in the citing sentence predictor.

Coarse Label: Background
Describing the citing sentences referring to the theory, principle, concept, topic, problem, etc. from cited papers. **Fine-grained Label:** • **(c0) definition,** explaining the definition of general theory, principle, concept, topic, problem, etc. *Example:* Neural Machine Translation (NMT) is a simple new architecture for translating texts from one language into another <citation>. • **(c1) suggest,** giving the reader a suggestion to refer, see more detail, and explore other cited papers. *Example*: The interested reader may dig deeper into this subject by referring to <citation>. • **(c2) judgment,** highlighting the positive/negative, useful/not-useful, etc. of concept, topic, problem, etc. *example*: The n-coalescent has some interesting statistical properties <citation>. • **(c3) technical,** explaining how a theory, principle, concept, topic, problem, etc. is applied. *Example*: The inference is done using blocked Gibbs sampling <citation>. • **(c4) trend,** explaining the significance of the research topic, theory, principle, concept, topic, problem, and etc. *example*: A recent trend <citation> challenge shows that deeper CNNs achieve better results.
Coarse Label: Citing Paper Work
What is proposed by the author? **Fine-grained Label:**

- **(c5) corroboration,** while proposing a research topic, citing paper cites cited paper. *Example*: We also briefly present a Minimum Message Length method <citation> of causal discovery in Section 4.
- **(c6) based on,** stating that citing paper follow, consider, is built based on, and inspired by the cited paper. *Example*: Instead, inspired by <citation>, we focus on the parallelism of the decoder and the energy consumed within it.
- **(c7) use,** citing paper use, implement, employ, or adopt the concept, dataset, technique, etc. *example*: We use the unsupervised dependency parser (UDP) implemented by <citation>.
- **(c8) extend,** citing paper extends, adapt, improves, adds, or modifies the cited paper' work. *Example*: Here we modify the microscopic search rules of <citation> to make it applicable to undirected graphs.
- **(c9) dominant,** the performance of citing paper outperforms cited paper' performance. *Example*: Note that our method outperforms the state of the art on both languages <citation>.
- **(c10) future,** mentioning the future plan of citing paper. *Example*: However, we will explore the distributed variants of the proposed S3GD like <citation> in the future.

Coarse Label: Cited Paper Work

What is done by cited papers.

Fine-grained Label:

- **(c11) propose,** describing the proposed research by the cited paper. *Example*: In <citation> the authors propose a model for storing and operating on infra-red images.
- **(c12) success,** highlighting the success of cited paper. *Example*: <citation> successfully extracts body appearance and topology from synthetic and real input.
- **(c13) weakness,** highlighting the weakness of cited paper. *Example*: The limitation of <citation> is that they only focused on two-user communication systems.
- **(c14) result,** describing the result of the cited paper (neutral). *Example*: The JavaBaker oracle has a precision of 0.97 and a recall of 0.83 <citation>.
- **(c15) dominant,** stating the superiority of cited paper compared to citing paper. *Example*: Only the deeper ResNet classifier <citation> outperformed our approach.

Coarse Label: Compare and Contrast

Compare and contrast is done between citing paper and cited paper.

Fine-grained Label:

- **(c16) compare,** describing the similarity between citing and cited papers. *Example*: The BLHT algorithm <citation> is closely related to our work.
- **(c17) contrast,** describing the differences between citing and cited papers. *Example*: However, unlike <citation>, our model does not have a partially nested information structure.

Coarse Label: Other

This label is prepared for citing sentences that do not match with the above criteria

Fine-grained Label:

- **(c18) comparison,** comparison between cited papers (whether similarities or differences between them). *Example*: Table compares the computational complexity of the proposed method with AOG <citation> and nCTE <citation>.
- **(c18) multiple_intent,** citing sentences have two or more citation marks for different intents. *Example*: One of the early works in the area of Property Testing is the work of Blum, Luby and Rubinfeld <citation> which dealt with linearity testing (see <citation> for low degree testing).
- **(c18) other,** this label is designed for citing sentences that do not meet all of the label categories described above. *Example*: The first paper is by Sab'an and Sethuraman <citation>.

4.1.2. Regular Sentence Predictor

The *regular sentence* predictor is the first additional predictor proposed in this thesis. This predictor is motivated by not all authors' reasons for making citations during manuscript writing can be accommodated using only *citing sentences*. Specifically, they provide detailed explanations after making citations. This predictor is designed by applying the scheme of *citation functions* to *regular sentences*. Accordingly, applying the scheme implies that all *regular sentences* extracted from each paper of the ICLR dataset are categorized using ML using the same model when classifying the *citing sentences*. Therefore, this predictor will have the same labels as the *citing sentence* predictor, and the labels are denoted as *r0 to r19*.

4.1.3. Reference-based Predictor

The second additional predictor proposed in this thesis is a *reference-based*. This predictor comprises 24 generic, preprint, and journal labels. These labels are generated by manually reviewing the reference section of the papers in the dataset. The reviewing process is in two aspects as follows: The first aspect involves checking well-known publications in both conferences and journals in AI, ML, Natural Language Processing, and Data Mining, among others; and the second aspect is appearing these publications in the reference section of the ICLR paper in the dataset. Additionally, the review shows that the papers are frequently cited in preprint repositories and references published within 3 years. Therefore, the labels are encoded as *ref0 to ref23* as shown in Table 4.4.

4.1.4. Combination Predictor

There is one more predictor comprising all the mentioned predictors. This combination predictor is proposed to examine whether the combined features of all predictors can generate optimum prediction performance compared with the features that belonged to a single predictor. The features in this predictor are denoted as *comb0 to comb63*.

Table 4.4. List of features in the reference-based predictor.

Generic Reference Labels
• **(ref0) NUM_REF**: Number of total references • **(ref1) NUM_REF_3YEARS**: Number of references within 3 years
Preprint Labels
• **(ref2) arXiv** Preprint Repository
Conference Venue Labels
• **(ref3) NeurIPS (formerly NIPS)**: Conference on Neural Information Processing Systems • **(ref4) ICLR**: International Conference on Learning Representations • **(ref5) ICML**: International Conference on Machine Learning • **(ref6) AAAI**: Association for the Advancement of Artificial Intelligence • **(ref7) ICCV**: International Conference on Computer Vision • **(ref8) CVPR**: Conference on Computer Vision and Pattern Recognition • **(ref9) EMNLP**: Empirical Methods in Natural Language Processing • **(ref10) ACL**: Association for Computational Linguistics • **(ref11) NAACL**: North American Chapter of the Association for Computational Linguistics • **(ref12) ECCV**: European Conference on Computer Vision • **(ref13) ICRA**: The International Conference on Robotics and Automation • **(ref14) ICASSP**: the International Conference on Acoustics, Speech, and Signal Processing • **(ref15) IJCAI**: The International Joint Conference on Artificial Intelligence • **(ref16) AISTATS**: The International Conference on Artificial Intelligence and Statistics • **(ref17) SIGKDD**: Special Interest Group on Knowledge Discovery and Data Mining
Journal Labels
• **(ref18) Neuralcom**: Neural Computation • **(ref19) IEEE Transaction** • **(ref20) ACM Transaction** • **(ref21) MIT Press** • **(ref22) Nature** • **(ref23) JMLR**: The Journal of Machine Learning Research

4.3. Building Prediction Features

This section discusses the prediction features for classification and regression tasks comprising several parts. Firstly, the beginning of this section describes the paper acceptance dataset as the primary data source employed in this thesis. Secondly, this section discusses the creation of prediction features and their distribution. Lastly, this section describes how the experiment scenarios are planned and executed.

4.3.1. The Dataset of Paper Acceptance

This thesis applies the dataset from (Yuan et al., 2022), which provided a well-parsed paper collection from the ICLR 2017–2020 and their equivalent final review decisions and review scores. The final review decision on whether the submitted papers are accepted or rejected is determined by the editor of the conference. The review scores are assigned by three reviewers ranging from 1 to 10, where the review score <4 is labeled as "rejected," that >7 is labeled as "accepted," and that of 5 and 6 are labeled as "marginally below" and "marginally above," respectively. These review scores are provided by the OpenReview[11] platform in the review process. Notably, the paper with marginal review scores can still be labeled as "accepted." Therefore, this study uses the average of three review scores from three reviewers to determine whether the paper is good or poor. A submitted paper can be labeled as poor when the average review score is ≤4 and good when the average review score is 4. The papers which had 4<average review scores<5 is assigned as the good category for several reasons. First, this score-range should be obtained from at least one reviewer who provides a review score of 5 or more; second, the paper in this category can be accepted by the editor; and third, the guide shows that scores of 4 or below will be rejected and no rule to reject the borderline scores of 5 and 6 directly. Since the review scores are the focus, this thesis does not consider whether the accepted paper will be presented as an oral, poster, or workshop. This thesis selected 5,156 papers out of 5,192 papers from the dataset. This difference occurs because it is could not determine the corresponding review results regarding the final review decisions or scores in many papers. Finally, the paper acceptance dataset for the final experimental comprises 1,722 and 3,434 accepted and rejected papers, respectively. This thesis also identified 3,575 and 1,581 good and poor papers, respectively, within the same dataset. Table 4.5 shows the detailed dataset distribution.

[11] https://openreview.net/

Table 4.5. Distribution of paper collection used in this thesis.

Year	Accepted Papers	Rejected Papers	Good Papers	Poor Papers	Total
2017	198	289	416	71	487
2018	336	571	769	138	907
2019	502	1,048	1,275	275	1,550
2020	686	1,526	1,115	1,097	2,212
Total	1,722	3,434	3,575	1,581	**5,156**

4.3.2. Building the Classification Features

The classification features are created by gathering each feature (label) of all predictors in the paper. Therefore, the proposed method extracts all *citing sentences, regular sentences,* and references from all papers in the dataset. For the first two predictors, i.e., *citing and regular sentences*, the extracted sentences are categorized into *fine-grained* labels using the developed ML model based on SciBERT (Beltagy et al., 2019) obtained from the previous stage (Basuki & Tsuchiya, 2022). Accordingly, the SciBERT model achieved an accuracy of 0.83, followed by an f1 score of 0.84. The prediction method applied the hyperparameters setting to obtain this performance as follows*: learning rate* $3e^{-5}$*, batch 32, class weight-based balanced* dataset. Notably the SciBERT was applied with the ktrain[12] python package. Conversely, for the *reference-based* predictor, the keyword matching approach is employed to estimate each label in all papers. Therefore, to create the combination predictor, the prediction method simply merges the features of all predictors to obtained 64 features (*atr0* to *atr63*). The final features will accompany the target label of *accepted-rejected* and *good-poor*.

4.3.3. Building the Regression Features

The review score prediction applies similar features as that in the classification tasks. The difference is that the review score prediction is considered a regression problem comprising two tasks, i.e., average and individual review score predictions. The average review score is

[12] https://github.com/amaiya/ktrain

obtained when the average review scores given by three reviewers are calculated. In contrast, each review score is given by each reviewer in the individual review score prediction. Here, the regression method treats the average and the individual review score predictions as single- and multi-output regressions, respectively. Therefore, both regression tasks will follow similar experiment settings.

4.3.4. The Distribution of Created Prediction Features

Therefore, this section presents the distribution of prediction features previously developed in the preceding section to provide a clear view of the proposed method used in this thesis.

Table 4.6. Distribution of each predictor in the dataset.

Paper Sources	Num. of Papers	Type of Classification Features	Num. of Instances
ICLR 2017	487	Citing Sentences	12,250
		Regular Sentences	72,940
		Reference-based	13,844
ICLR 2018	907	Citing Sentences	24,238
		Regular Sentences	141,741
		Reference-based	27,670
ICLR 2019	1,550	Citing Sentences	43,690
		Regular Sentences	254,961
		Reference-based	52,838
ICLR 2020	2,212	Citing Sentences	66,651
		Regular Sentences	383,915
		Reference-based	82,019

Here, the instance distribution of all predictors is discussed. Table 4.6 shows the yearly distribution. Figure 4.2 depicts the distribution of entire years. In Figure 4.2.1 and Figure 4.2.2, it is clearly observed that labels in the *citing sentence* predictor significantly vary compared with the *regular sentence* predictor. This trend is caused by using labels in the *regular sentence* predictor adopted from the *citing sentence*. In Figure 4.2.3, the spread of labels in the *reference-based* predictor is dominated by the number of references for the last 3 years *(NUMREF2YEARS)*, followed by preprint source *(arXiv)*, *ICLR*, *NeurIPS*, and *ICML*. Furthermore, the other labels in this predictor possess relatively equal distribution.

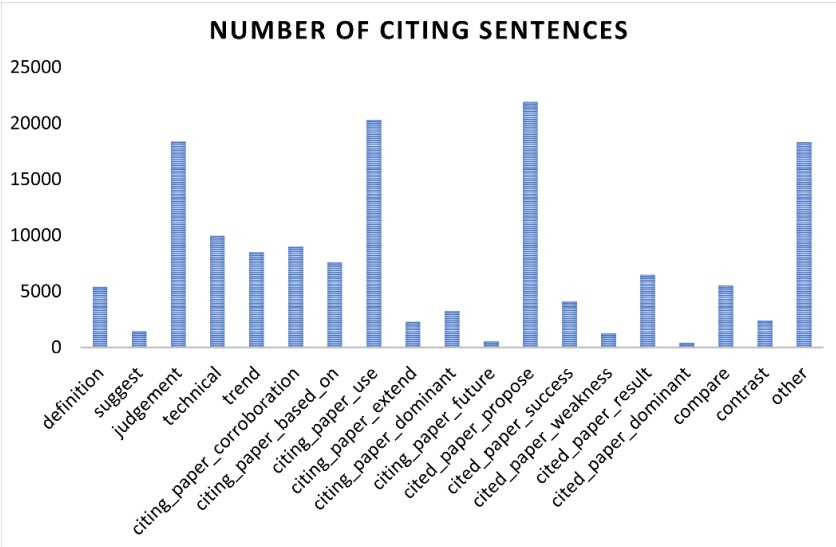


Figure 4.2.1 – Distribution of citing sentences.

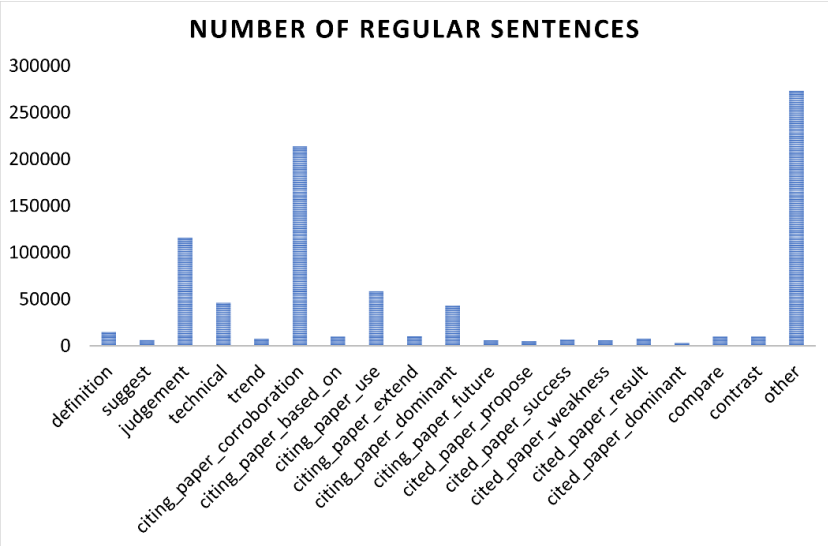


Figure 4.2.2 – Distribution of regular sentences.

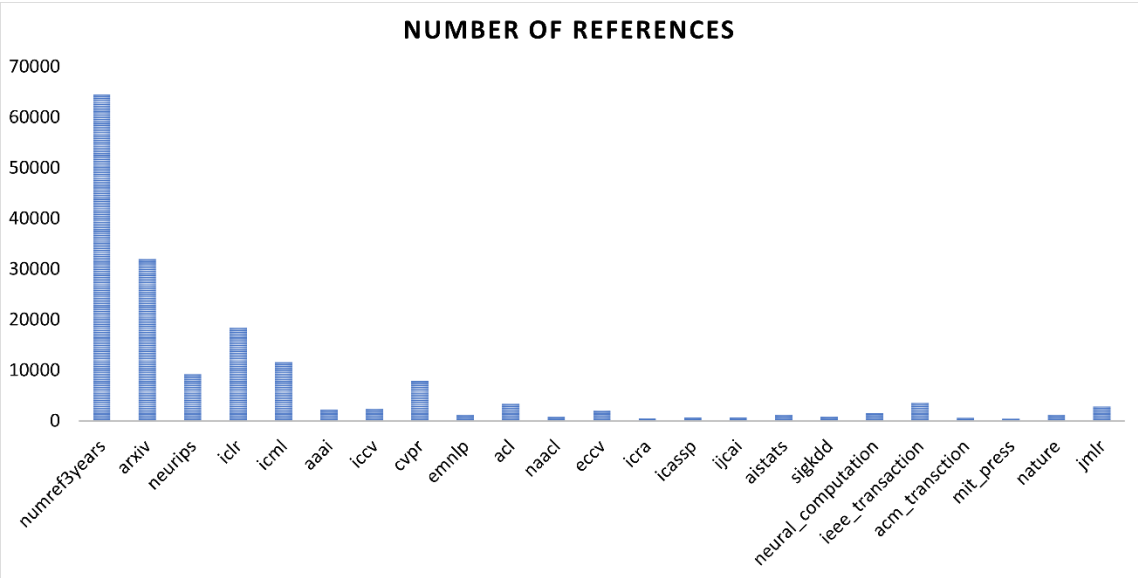


Figure 4.2.3 – Distribution of references.

Figure 4.2. The distribution of all classification features in ICLR from 2017 to 2020 is presented on each attribute.

** Figure 2.3 does not present the number of reference distributions (NUM_REF) because it is obtained by accumulating all other labels' distributions.*

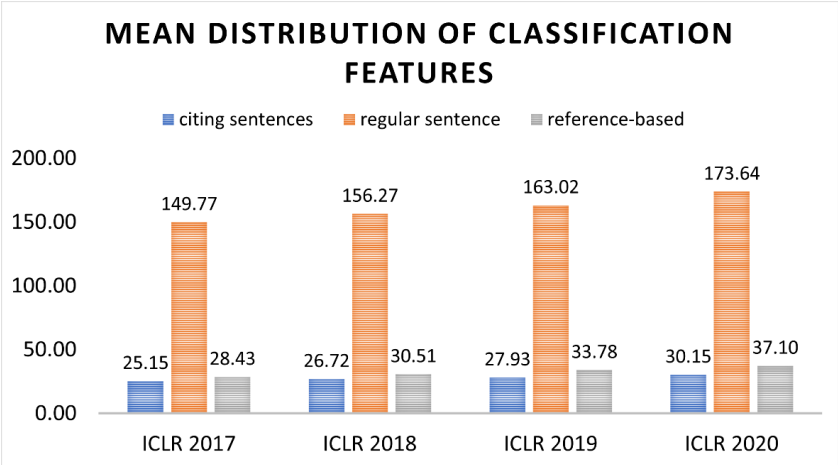


Figure 4.3. The means' distribution of the citing and regular sentences in the ICLR datasets.

** Here, the x and y axes represent the sentences' categories and means, respectively.*

4.3.5. Experiment Scenario

Here, the *accepted-rejected* and *good-poor* predictions are treated as classification issues. Both prediction tasks apply similar experimental settings as follows: there are four experiment scenarios, with each scenario representing each type of predictor. Specifically, the experiment on the *citing sentence*, *regular sentence*, *reference-based*, and combination predictors adopt features c0 to c19, r0 to r19, ref0 to ref23, and comb0 to comb63, respectively. XGBoost is used as a ML algorithm for all experiments and three FS methods to show the most influential features. Additionally, the FS methods employed here are Chi-square (Chi2), Recursive Feature Elimination (RFE), and Sequential Feature Selector (SFS) Forward. Notably, the FS methods are implemented using the python scikit-learn library[13]. The FS method experiment is conducted by observing the classification performances based on the number of selected features, beginning from a single feature to the maximum number of features. Therefore, the prediction method evaluates the data balancing technique's impact on the classification performances using Synthetic Minority Over-sampling Technique (SMOTE)-based method.

Conversely, this thesis proposes using five regression algorithms and one FS method in the regression experiment. The regression algorithms used here are the Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Support Vector Regression (SVR), Extreme Gradient Boosting Regression (XGBR), and Decision Tree Regression (DTR). Alternatively, the FS method used here is SelectKBest, based on the python library. In each experiment, the FS observes the regression performance starting from a single feature to the maximum number of features. Therefore, this study uses MAE and RMSE as performance metrics. Notably, all the regression algorithms and FS method are implemented using the scikit-learn python library.

[13] https://scikit-learn.org/stable/

4.4. Prediction Experiment Results

This section describes the experiment results for predicting paper quality, which is classified into three parts, i.e., the results of the *accepted-rejected*, the *good-poor*, and the review scores tasks, respectively. Furthermore, the results cover prediction performances measured by several metrics and the most influential features to achieve the best performances. Moreover, this section also provides an analysis of the performances against the real review scores, the phenomenon of meaning shifts of *regular sentence* predictors, and the performance comparison between this thesis and previous studies.

4.4.1. Performance of Classification Tasks

Tables 4.7 and 4.8 present the best results of all scenarios in *accepted-rejected* and *good-poor* tasks, respectively. Therefore, this study uses additional metrics such as precision, recall, AUC, and f1 for two reasons instead of using only a single accuracy metric. First, the accuracy can be biased toward most classes in an imbalanced setting. Second, recall by setting accepted or good papers as a positive label should be a more suitable metric in this study. This result is because predicting as many positive instances as possible is better than wrongly predicting positive instances into negative classes.

In the *accepted-rejected* task, the best accuracy was 0.73, which was achieved using the combination feature, SFS Forward, and 15 features in the balanced setting. This scenario was also considered the best setting since it achieved 0.50 recall (second best result), 0.61 precision (best result), 0.72 AUC (one of the best results), and 0.55 f1 (one of the best results). Another remarkable result is that the same accuracy of 0.71 was obtained by applying a combination feature with two FS approaches, such as Chi2 and RFE, in the balanced setting. In the imbalanced setting, the *reference-based* and combination features had accuracies of 0.71 and 0.70, respectively, which were slightly lower than the best result in the balanced setting.

Generally, the imbalanced setting generated lower performance in all metrics than the balanced setting. The proposed classification approaches are less effective for determining the paper acceptance ratio even if it reached reasonable accuracies of more than 0.70 considering the entire performance.

In the *good-poor* tasks, the highest accuracies were 0.75 achieved using a combination of balanced settings, combination features, and three FS methods, such asChi2 (55 features), SFS Forward (using 45 features), or RFE (using 21 features). Although all FS methods in this setting showed similar accuracies, the Chi2 was slightly better than the others by showing a recall of 0.94. Furthermore, focusing on the imbalanced setting, the achieved accuracy of 0.74 was slightly lower than in the balanced setting. However, all performance metrics in the imbalanced setting generally revealed better results than those in the balanced setting. For example, the minimum accuracy, recall, and f1 in the imbalance setting are 0.72, 0.92, and 0.82, respectively, while in the balanced setting are 0.62, 0.66, and 0.71, respectively. Additionally, the imbalanced setting required less than 10 features for most settings and only a single feature (using Chi2 applied to referenced-based and combination types of features) to achieve reasonable accuracies of 0.72 in several settings.

Focusing on **obtaining as many positive instances as possible** through recall can provide broader performance measurements. The best recall on the imbalanced and balanced settings showed 0.37 and 0.63, respectively, which were considered ineffective for **the *accepted-rejected* task**. On the ***good-poor*** task, the recalls obtained the highest results by 0.99 using *citing sentence* predictors with all FS methods in the imbalanced setting. Interestingly, this recall was achieved using less than 10 features as follows: 8 features (Chi2), 8 features (SFS Forward), and 7 features (RFE). Conversely, in the balanced setting, the best recall was 0.94, achieved using the combination feature and Chi2. Notably, the balanced setting exhibited its consistency in applying the identical experiment configuration resulting in the best results

based on accuracy and recall. All the performances proved that the *citation functions* are quite representative in predicting the quality of the manuscript, whether good or poor.

Table 4.7. Best performances of each scenario in the accepted-rejected prediction.

Imbalanced Setting							
Predictor	**FS Methods**	**Num. of Features**	**Accuracy**	**Recall**	**Precision**	**AUC**	**F1**
Citing Sentence	Chi2	2	0.67	0.15	0.54	0.63	0.24
	SFS Forward	1	0.67	0.02	1.00	0.57	0.03
	RFE	6	0.66	0.12	0.48	0.62	0.19
Regular Sentence	Chi2	4	0.66	0.14	0.49	0.60	0.22
	SFS Forward	1	0.68	0.05	0.73	0.61	0.09
	RFE	7	0.67	0.16	0.51	0.61	0.24
Reference Based	Chi2	19	0.70	0.28	0.62	0.67	0.39
	SFS Forward	13	**0.71**	0.24	**0.69**	0.68	0.36
	RFE	12	0.70	0.26	0.63	0.66	0.36
Combination	Chi2	13	**0.71**	0.33	0.62	0.70	0.43
	SFS Forward	37	**0.71**	**0.37**	0.60	**0.73**	**0.46**
	RFE	18	0.70	0.36	0.59	0.70	0.45
Balanced Setting							
Predictor	**FS Methods**	**Num. of Features**	**Accuracy**	**Recall**	**Precision**	**AUC**	**F1**
Citing Sentence	Chi2	19	0.66	0.27	0.47	0.65	0.35
	SFS Forward	18	0.66	0.32	0.47	0.67	0.38
	RFE	16	0.66	0.31	0.50	0.65	0.38
Regular Sentence	Chi2	15	0.67	0.29	0.51	0.64	0.37
	SFS Forward	17	0.67	0.25	0.52	0.64	0.34
	RFE	11	0.67	0.31	0.51	0.63	0.39
Reference Based	Chi2	24	0.64	0.61	0.47	0.67	0.53
	SFS Forward	2	0.65	0.10	0.38	0.51	0.16
	RFE	21	0.66	**0.63**	0.49	0.68	**0.55**
Combination	Chi2	28	0.71	0.50	0.57	**0.72**	0.53
	SFS Forward	15	**0.73**	0.50	**0.61**	**0.72**	**0.55**
	RFE	58	0.71	0.49	0.57	**0.72**	0.53

The impact of *citation functions* in the classification tasks is analyzed through the following two aspects: the classification performances and the number of features to achieve the best performance. The impact of *citation functions* is more dominant in the *good-poor* task than the *accepted-rejected* task, particularly in the imbalanced scenario. For example, the best recalls were obtained using the *citation functions*-based prediction by 0.99 (*citing sentences* predictor)

and 0.98 (*regular sentences* predictor). As mentioned above, attaining as much high recall as possible is important to get as many good papers as possible, which is more reasonable and applicable for assisting the editor in *filtering* the submitted manuscripts. Additionally, this highest recall was obtained by employing the fewest number of features by 7 when combining the *citing sentences* predictor with the RFE.

Table 4.8. Best performances of each scenario in the good-poor prediction.

Imbalanced Setting							
Predictor	**FS Methods**	**Num. of Features**	**Accuracy**	**Recall**	**Precision**	**AUC**	**F1**
Citing Sentence	Chi2	8	0.72	**0.99**	0.72	0.65	0.83
	SFS Forward	8	0.72	**0.99**	0.72	0.65	0.83
	RFE	7	0.72	**0.99**	0.71	0.62	0.83
Regular Sentence	Chi2	10	0.72	0.98	0.72	0.62	0.83
	SFS Forward	17	0.72	0.98	0.72	0.63	0.83
	RFE	17	0.73	0.98	0.73	0.61	0.83
Reference Based	Chi2	1	0.72	0.92	**0.74**	0.68	0.82
	SFS Forward	17	0.73	0.95	**0.74**	0.70	0.83
	RFE	17	0.72	0.94	0.73	0.71	0.83
Combination	Chi2	1	0.72	0.92	**0.74**	0.68	0.82
	SFS Forward	59	**0.74**	0.96	**0.74**	0.70	**0.84**
	RFE	32	**0.74**	0.96	**0.74**	**0.72**	**0.84**
Balanced Setting							
Predictor	**FS Methods**	**Num. of Features**	**Accuracy**	**Recall**	**Precision**	**AUC**	**F1**
Citing Sentence	Chi2	18	0.67	0.73	0.78	0.65	0.76
	SFS Forward	19	0.66	0.72	0.77	0.66	0.75
	RFE	18	0.67	0.73	0.77	0.66	0.75
Regular Sentence	Chi2	13	0.64	0.66	0.78	0.62	0.71
	SFS Forward	11	0.63	0.66	0.78	0.63	0.71
	RFE	20	0.62	0.66	0.77	0.63	0.71
Reference Based	Chi2	10	0.66	0.66	**0.81**	0.68	0.73
	SFS Forward	24	0.65	0.70	0.78	0.65	0.74
	RFE	18	0.66	0.68	0.80	0.66	0.73
Combination	Chi2	55	**0.75**	**0.94**	0.76	**0.73**	**0.84**
	SFS Forward	45	**0.75**	0.93	0.76	**0.73**	**0.84**
	RFE	21	**0.75**	0.92	0.76	0.71	0.83

Analysis of the Most Important Features of Classification Experiments

This section reports the analysis of the selected features obtained using the FS methods, particularly the top 10 most important features adopted by the *combination* predictor (this predictor achieved the best performances in both prediction tasks). The most important features presented here encompass both imbalanced and balanced settings, with 60 selected features in each prediction task. Tables 4.9 and 4.10 show the distribution of selected features categorized based on predictors and *coarse* labels of *citation functions*, respectively. The distribution of these two tables is obtained from Table 4.11, and Table 4.12 shows the detailed selected features in the *accepted-rejected* and *good-poor* tasks, respectively.

Table 4.9. Distribution of the top 10 most important features categorized based on the predictors.

predictor	Frequencies	
	accepted-rejected task	good-poor task
citing sentence	12	14
regular sentence	28	26
reference-based	20	20

Notably, the top 10 most important features were dominated by features belonging to the *regular sentence* predictor, indicating the highest frequency of 28 and 26 in the *accepted-rejected* and *good-poor* tasks, respectively. These results are strongly influenced because this predictor has the highest number of instances compared with other predictors (see Table 5). The second highest frequency was obtained by features belonging to the *reference-based* predictor by signifying a frequency of 20 in both prediction tasks. The *citing sentence* predictor has the lowest frequency by 12 and 14 in the *accepted-rejected* and *good-poor* tasks, respectively.

This study reports other notable findings, further investigating the top 10 most important features. The significant highest frequency is shown by *fine-grained* features belonging to *citing paper work* by 17 and 14 in the *accepted-rejected* and *good-poor* tasks, respectively. These significant *fine-grained* features were *citing_paper_use*, *citing_paper_future*,

citing_paper_dominant, and *citing_paper_corroboration*. The second highest frequency was the *number of citing sentences* or *number of regular sentences*, with 8 and 12 in the *accepted-rejected* and *good-poor* tasks, respectively. A slightly lower distribution is shown by *background* by 7 and 8 in the *accepted-rejected* and *good-poor* tasks, respectively. Although *fine-grained* features belonging to *cited paper* have only a few frequencies, that related to the *compare and contrast* showed zero frequency. The zero frequency in the *compare and contrast* is caused by low instance distribution in the dataset. Notably, the *citing_paper_dominant* had high frequencies, although it has few instances distributions in the dataset (see Figure 4.2.1 and Figure 4.2.2).

Table 4.10. Distribution of the top 10 most important features categorized based on the coarse labels.

Feature coarse categories	Frequencies	
	Accepted-rejected task	Good-poor task
Background	7	8
Citing paper work	17	14
Cited paper work	4	1
Compare and contrast	0	0
Other	4	5
Number of citing sentence & regular sentences	8	12
Generic reference	8	10
Preprint	0	0
Conference venue	8	6
Journal	4	4

NOTE:
The *coarse* labels, i.e., *background, citing paper work, cited paper work, compare and contrast,* and *other,* are representation of the *citing sentence* predictor and the *regular sentence* predictor. The *coarse* labels falling into reference-based predictors are *generic reference, preprint, conference venue,* and *journal*.

Identifying the features based on the *reference-based* predictor depicted that the highest frequencies are obtained by a generic reference containing two features, i.e., *num_ref* and *num_ref_3years*, by showing values of 8 and 10 in the *accepted-rejected* and the *good-poor* task, respectively. The features belonging to the conference venue show the small lower frequencies by showing the distribution of 8 and 6 in the *accepted-rejected* and *good-poor* tasks, respectively. The journal venue showed few frequencies of 4 in both prediction tasks;

however, the preprint (arXiv) revealed the zero-frequency but had significant instance distribution in the dataset (see Figure 4.2.3).

Another fascinating finding in the conducted experiments is that the *citation functions*-based predictors (*citing* and *regular sentence* predictors) are more influential than the *reference-based* predictor. Two experiment results support this fact. First, the distribution of features belonging to the *regular sentences* predictor has the highest number in the experiment using a combination predictor in both prediction tasks (Table 4.9). This trend implies that this predictor contributes more to the prediction results. Second, using a few features, the *citing sentences* predictor obtained the highest recall in the *good-poor* task. Additionally, this highest result is one of the most important findings since obtaining as many good papers as possible is crucial in the review process. Finally, although the *reference-based* predictor, when considered, reached slightly higher accuracy in the *accepted-rejected* task when using the imbalanced setting, the balanced setting for the same task or both imbalanced and balanced settings on *good-poor* task had accuracy reaching the same or even lower results compared with *citation functions*-based predictors. Altogether, the *reference-based* predictor still contributes to forming the combination predictor, although the *citation functions*-based predictors have more impact in obtaining the best results.

Analysis toward the Real Review Scores of Classification Experiments

It is worth discussing why the prediction models were effective in the *good-poor* task rather than the *accepted-rejected* task. Accordingly, the review scores of ICLR 2017–2020 are depicted in Figure 4.4 and the mean and variance of review scores of the best results in both classification tasks in Figure 4.5. The boundaries between TP (True Positive) versus TN (True Negative) and FP (False Positive) versus FN (False Negative) in the mean of review scores are clearly separated in the *good-poor* task but unclear in the *accepted-rejected* task. However, the two classification tasks show a similar pattern in the distribution of variances. The only

prominent difference is that TP has the most paper in the *good-poor* tasks, whereas TN has the highest number in the *accepted-rejected* task. This variation occurs because the achieved recall on the *good-poor* task is greater than in the *accepted-rejected* task. Summarily, the proposed classification features are more effective at categorizing whether the paper is good or poor rather than predicting its acceptance rate.

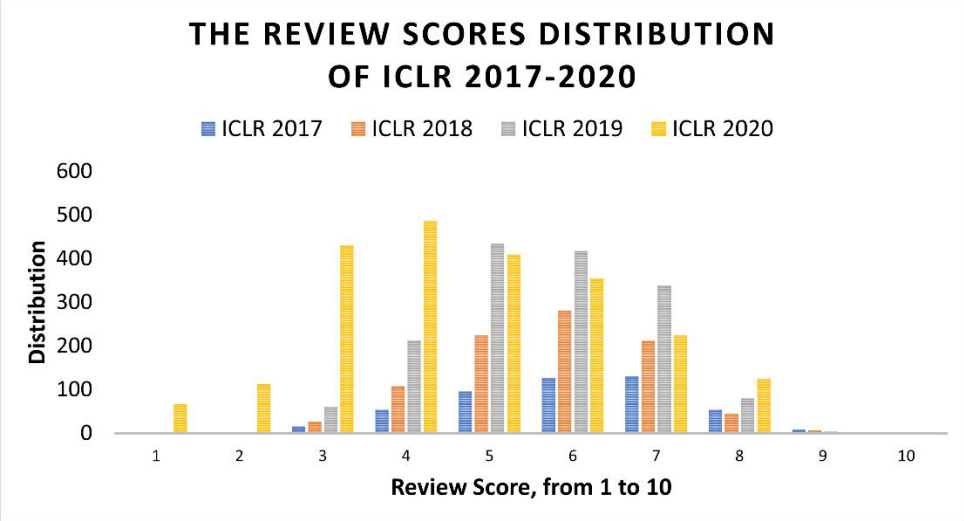


Figure 4.4. Distribution of review scores in ICLR from 2017 to 2020.

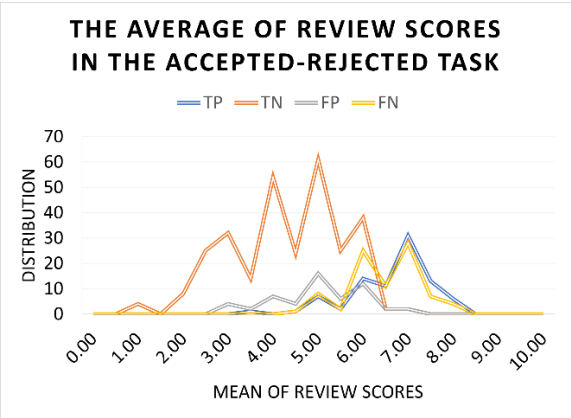


Figure 4.5.1 – The average review scores in accepted-rejected task

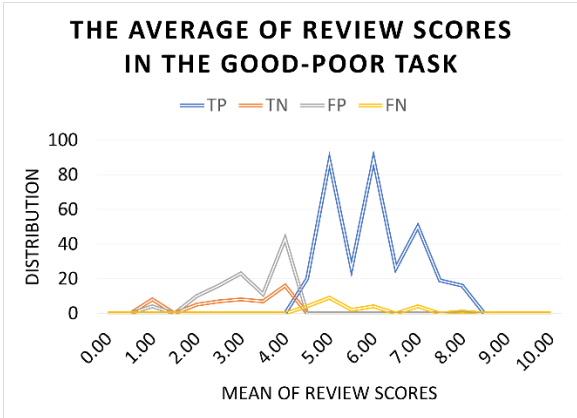


Figure 4.5.2 – The average review scores in good-poor task

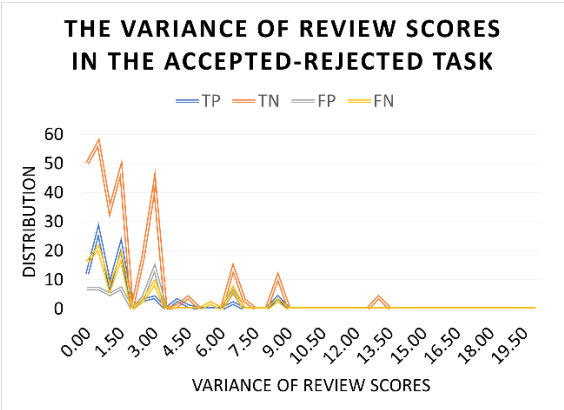


Figure 4.5.3 – The variance of review score in accepted-rejected task

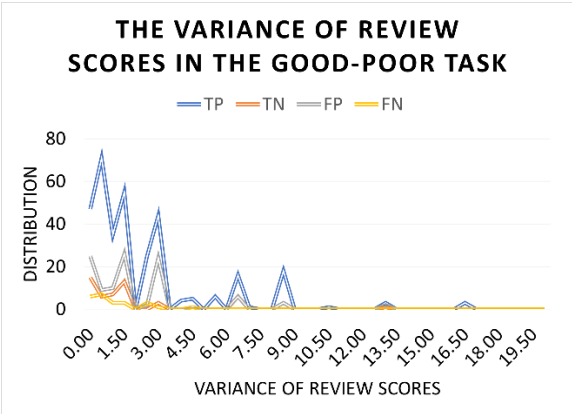


Figure 4.5.4 – The variance of review score in good-poor task

Figure 4.5. Distribution of mean of review scores and variance of review score of the best results in accepted-rejected and good-poor tasks.

The Meaning Shift of Regular Sentence Predictor

Since the *citing sentence* predictor's attributes are designed for *citing sentences*, they must be checked for compliance with *regular sentences*. The compliance check is performed by randomly selecting 1,000 samples from labeled sentences and evaluating the label for each sentence. This procedure reveals that, while several labels' meanings shifted, other labels remain relevant with the original definition adopted from the *citing sentence*. This occurred because the ML models struggle to recognize clear indications of whether a *regular sentence* describes a *citing paper* or *cited paper*. For example, the *coarse* label *background* does not experience the meaning shift compared with other *coarse* label *compare and contrast*, which mainly discusses the similarity and difference between *citing paper* and *cited paper*. Although several attributes' meanings shifted, they still retained the same idea as the original attributes. Table 4.13 presents a detailed explanation of this phenomenon.

Table 4.11. The top 10 most important features of the combination predictor in the accepted-rejected prediction.

Rank	Imbalance Setting			Balanced Setting		
	Chi2	**SFS Forward**	**RFE**	**Chi2**	**SFS Forward**	**RFE**
1st	#2 - number of regular sentences	#3 - ijcai	#2 - number of regular sentences	#2 - number of regular sentences	#3 - ijcai	#2 - number of regular sentences
2nd	#2 - citing_paper_corroboration	#3 - acm_tran	#2 - other	#2 - citing_paper_corroboration	#3 - acm_tran	#2 - other
3rd	#2 - other	#1 - suggest	#2 - citing_paper_corroboration	#2 - other	#1 - suggest	#2 - citing_paper_corroboration
4th	#3 - num_ref_3years	#3 - aistats	#1 - number of citing sentences	#3 - num_ref_3years	#3 - aistats	#1 - number of citing sentences
5th	#1 - number of citing sentences	#3 - mit_press	#3 - num_ref_3years	#1 - number of citing sentences	#3 - mit_press	#3 - num_ref_3years
6th	#3 - num_ref	#1 - citing_paper_future	#3 - num_ref	#3 - num_ref	#1 - citing_paper_future	#3 - num_ref
7th	#2 - citing_paper_use	#1 - number of citing sentences	#2 - judgment	#2 - citing_paper_use	#1 - number of citing sentences	#2 - judgment
8th	#3 - neurips	#2 - number of regular sentences	#2 - citing_paper_use	#3 - neurips	#2 - number of regular sentences	#2 - citing_paper_use
9th	#2 - judgment	#3 - num_ref_3years	#2 - technical	#2 - judgment	#3 - num_ref_3years	#2 - technical
10th	#3 - citing_paper_dominant	#1 - other	#2 - citing_paper_dominant	#2 - citing_paper_dominant	#1 - cited_paper_propose	#2 - citing_paper_dominant

Table 4.12. The top 10 most important features of the combination predictor in the good-poor prediction.

Rank	Imbalance Setting			Balanced Setting		
	Chi2	**SFS Forward**	**RFE**	**Chi2**	**SFS Forward**	**RFE**
1st	#2 - number of regular sentences	#3 - naacl	#2 - number of regular sentences	#2 - number of regular sentences	#3 - naacl	#2 - number of regular sentences
2nd	#3 - num_ref_3years	#1 - citing_paper_future	#2 - other	#3 - num_ref_3years	#1 - citing_paper_future	#2 - other
3rd	#2 - citing_paper_corroboration	#3 - mit_press	#2 - citing_paper_corroboration	#2 - citing_paper_corroboration	#3 - mit_press	#2 - citing_paper_corroboration
4th	#2 - other	#3 - eccv	#1 - number of citing sentences	#2 - other	#3 - eccv	#1 - number of citing sentences
5th	#1 - number of citing sentences	#3 - ijcai	#3 - num_ref_3years	#1 - number of citing sentences	#3 - ijcai	#3 - num_ref_3years
6th	#3 - num_ref	#1 - cited_paper_dominant	#3 - num_ref	#3 - num_ref	#1 - cited_paper_dominant	#3 - num_ref
7th	#2 - citing_paper_use	#3 - acm_tran	#2 - judgment	#2 - citing_paper_use	#3 - acm_tran	#2 - judgment
8th	#2 - citing_paper_dominant	#1 - suggest	#2 - citing_paper_use	#2 - citing_paper_dominant	#1 - suggest	#2 - citing_paper_use
9th	#2 - judgment	#1 - cited_paper_weakness	#2 - citing_paper_dominant	#2 - judgment	#1 - cited_paper_weakness	#2 - citing_paper_dominant
10th	#1 - citing_paper_use	#3 - icra	#2 - citing_paper_use	#1 - citing_paper_use	#3 - icra	#2 - technical

Note: For explaining the top 10 of the most important features in Table 4.11 and Table 4.12, the mark # is used to denote the type of predictors. For instance, #1: citing sentence predictor, #2: regular sentence predictor, and #3: reference-based predictor. Note that predictor #1 and predictor #2 use the same feature name because predictor #2 is created based on predictor #1.

Table 4.13. The meaning shifts explanation of fine-grained labels.

Coarse Label: Background
In this coarse class, there are four labels that exactly match the original definitions and a label needs to be adjusted. **Fine-grained Label:** • **(atr0) definition,** *meaning shift:* no; *example*: Jensen-Shannon divergence is a smoothed symmetric variant of KL divergence. • **(atr1) suggest,** *meaning shift*: yes, by suggesting an internet source, refer other sections in the paper, etc.; *example*: Additional results on the effects of distributed training on representation drift and Q-value discrepancy are given in the Appendix. • **(atr2) judgment,** *meaning shift*: no; *example*: Such high inference cost is near-infeasible in many online and latency-critical applications. • **(atr3) technical,** *meaning shift*: no; *example*: The codebook is then learned by minimizing the following objective function. • **(atr4) trend,** *meaning shift*: no; *example*: There are a number of existing solutions to both of these challenges, but they fall short.
Coarse Label: Citing Paper Work
Four labels need to expand their definition, two labels show identical definition with citing sentence.
Fine-grained Label: • **(atr5) corroboration,** *meaning shift*: yes, with discussing the contribution of the proposed research, including how citing papers apply a certain concept/method/etc.; *example*: In this thesis D is taken as KullbackLeibler divergence (dkl) to measure the similarity of policies. • **(atr6) based on,** *meaning shift*: no; *example*: Our work follows the most reliable and widely used robust model approach ,Äî adversarial training, which finds a set parameter to make the model robust. • **(atr7) use,** *meaning shift*: yes, citing study uses certain methods without specifying the source of such methods (no indication action/implement/apply indication); *example*: We choose to use EB since it produces a valid probability distribution for each network layer. • **(atr8) extend,** *meaning shift*: yes, by stating that citing paper extend/improve/update/etc. certain techniques; *example*: no example. • **(atr9) dominant,** *meaning shift*: yes, which explains the success of citing paper, and sometimes, explaining the result of citing paper; *example*: Table shows that our bounds are a super set to true bounds computed with an exact MIP solver. • **(atr10) future,** *meaning shift*: no; *example*: One of our future extensions is to adapt the current model to predict more dynamic outputs.
Coarse Label: Cited Paper Work
Almost all labels in Cited Paper Work are difficult to apply on the regular sentences. This is because it is difficult to find regular sentences explaining previous studies.
Fine-grained Label: • **(atr11) propose,** *meaning shift* **yes**, by stating the purpose of a certain techniques; *example*: The Transformer an attention-based neural network was introduced to improve machine translation and transduction. • **(atr12) success,** *meaning shift*: **yes**, by stating the success of certain techniques; *example*: Ablation studies show that improvements can be attributed to the use of TPRs in both the encoder and decoder to explicitly capture relational structure to support reasoning. • **(atr13) weakness,** *meaning shift*: **yes**, by mentioning the drawbacks of certain techniques; *example*: Another drawback for GPs is that it cannot handle graph-data directly without a special encoding scheme. • **(atr14) result,** *meaning shift*: **yes**, using explaining the results produced by certain techniques; *example*: ResNet35 results for randomly split sets of target objects on 4 environments from the replica dataset. • **(atr15) dominant** *meaning shift* difficult to adjust; *example*: no example.
Coarse Label: Compare and Contrast
In this coarse class, both compare and contrast must expand its definition.
Fine-grained Label:

• **(atr16) compare,** *meaning shift*: yes, stating the similarity between method used in citing paper with certain methods; *example*: Similar to the supervised learning setting we use current meta-parameters to optimize policy parameters under the current dynamics model. • **(atr17) contrast,** *meaning shift*: yes, by stating the reason why citing paper uses a specific technique instead of another; *example*: This is why we concentrate on ImageNet as opposed to MNIST or CIFAR.
Coarse Label: Other
No need adjustment.

Performance Comparison of Classification Experiments in this thesis with Previous Works

Generally, several existing works used accuracy as the only performance metric. Two studies employed alternative metrics, such as (Vincent-Lamarre & Larivière, 2021) using the f1, and (Bao et al., 2021) which employed the AUC. The other three studies employed more than one metric such as (Skorikov & Momen, 2020) which used accuracy, precision, recall, and f1, (Maillette de Buy Wenniger et al., 2020) which used accuracy and AUC, and (Ribeiro et al., 2021) which used accuracy, recall, and f1. Here, this thesis applied five metrics, i.e., accuracy, precision, recall, f1, and AUC (see Tables 4.7 and 4.8). Table 14 shows the detailed comparison.

Table 4.14. The accuracy-focused performance comparison between this thesis and previous works.

Existing Works	Dataset	The Best Accuracy
(Kang et al., 2018)	arXiv.CS	0.79
(K. Wang & Wan, 2018)	ICLR 2017	0.78#
(Jen & Chen, 2018)	ICLR 2017	0.71
(Ghosal et al., 2019)	ICLR 2017–2018	0.71#
(Ghosh et al., 2020)	ICLR 2017	0.65
(Skorikov & Momen, 2020)	arXiv and ICLR	0.83
(Maillette de Buy Wenniger et al., 2020)	arXiv.CS.CL	0.81
(Ciloglu & Merdan, 2020)	arXiv and ICLR	0.78#
(Joshi et al., 2021)	ICLR 2017	0.85
(Ribeiro et al., 2021)	AI conference, Robotics	0.77#
(Fytas et al., 2021)	ICLR 2017	0.88#
(Bharti et al., 2021)	ICLR 2019	0.77#
(Pradhan et al., 2021)	ICLR 2019	0.85#
This work: accepted-rejected task	ICLR 2017–2020	**0.73**
This work: good-poor task	ICLR 2017–2020	**0.75**

NOTE:
Mark (#) indicates that the research employed part of review comments or review scores as prediction features. The accuracies reported in this table represent the best performance achieved by each paper against a specific dataset. Readers may refer to each work for the complete performance of each work. This study noted that studies by (Vincent-Lamarre & Larivière, 2021) and (Bao et al., 2021) used f1 and AUC, respectively. Since the

best results of these works include the arXiv dataset obtained from (Kang et al., 2018), which determined the acceptance status as accepted and "probably-rejected," this thesis prefer not to put the results in this Table. Moreover, the comparison is also inapplicable to the study by (A. J. Casey et al., 2019) because they only focused on the Related Work section, annotators rather than the editor determined the target classes.

The best performance was achieved by (Joshi et al., 2021) showing an accuracy of 0.85 on a relatively small ICLR 2017 dataset. However, these results have some limitations as follows: no other metrics were used to show the performances under imbalanced situations. Second, accuracy was biased toward most classes. Third, since this work applied pre-defined (handcrafted) features, the results are less insightful for helping the peer-review process. Other promising results were (Skorikov & Momen, 2020) and (Maillette de Buy Wenniger et al., 2020) which achieved accuracies of 0.83 and 0.81, respectively. These two works used the arXiv dataset proposed by (Kang et al., 2018) that the papers' acceptance in the dataset were determined using two labels, i.e., accepted or "probably-rejected." Therefore, an issue regarding the confident level of the achieved accuracies existed. Several works obtained other competitive results by showing accuracies of more than 0.75. However, most of these studies used part of the review results as classification features. This approach is considered unfair since the acceptance prediction should be based on the manuscript. The prediction method proposed in this thesis achieved accuracy of 0.73. Therefore, considering the abovementioned issues, this result was competitive since the model in this thesis was developed using 15 classification features from the paper manuscript. Another perspective of the paper quality showed that the *good-poor* task achieved 0.75 of the best accuracy, which is considered slightly better than the best accuracy achieved in this thesis in the *accepted-rejected* task. However, the *good-poor* task obtained a high recall of 0.94 and competitive f1 of 0.84 using the same experimental setting.

Another interesting comparison can be obtained between this thesis study and that of (A. J. Casey et al., 2019) in which this thesis has developed a predictor containing a labeling scheme of the author's intentions to predict the paper quality. The difference is that while their work used the author's intentions in the Related Work section, which may cover both citing and *regular sentences*, this thesis used the author's intentions through *citation functions* represented

by citing sentences in the entire paper. Although the comparison cannot be performed directly because of the difference in the dataset and the target classes, this thesis showed that the labeling scheme of *citation functions* (*citing sentence* predictor) used here achieved better results in the *good-poor* task by showing the best accuracy and recall of 0.72 and 0.99, respectively. However, note that (A. J. Casey et al., 2019) showed the best accuracy of 0.7 in the poor-average-good task. These findings indicate that the labeling scheme of *citation functions* proposed in this thesis is more effective than the intention labels proposed in Casey's work. Additionally, covering the author's intention in the entire section of this thesis is crucial to assess the paper's quality rather than only in the Related Work section.

4.4.2. Performance of Regression Tasks

This section presents the regression task experiment results for predicting the average review score (Table 4.15), the individual review score (Table 4.16), and the top 10 most influential features in both regression tasks (Table 4.17).

The experiments show that the combination predictor achieved the best performances in both regression tasks by showing the lowest RMSE and MAE results. For example, in the average review score prediction, the lowest RMSE was 1.34, which RFR, GBR, and XGBR reached. Conversely, RFR and XGBR achieved the MAE's lowest results by demonstrating 1.07 points. DTR's best results required only a single feature in this regression task.

Conversely, the overall performances were worse in the individual review score prediction than the performance in the average review score prediction. The best results in the individual review score prediction was 1.71 for RMSE and 1.38 for MAE. Additionally, these results were produced by incorporating the combination predictor with RFR for RMSE and SVR for MAE. Interestingly, all best performances demonstrated by DTR require only a single feature, as in the average review score prediction task.

The impact of a predictor on the regression performances can be explained by comparing the performances (RMSE, MAE) and the number of features needed to obtain the best results. The *citation functions*-based predictors (*citing sentence* and *regular sentence* predictors) obtained slightly lower performances than the *reference-based* and the combination predictor in both the average and individual score prediction. However, the *citation functions*-based predictors require lesser features to achieve the best performances.

It is worth noting that the features representing the number of instances belonging to each feature or predictor were the most important in each predictor. For example, the rank-1 feature was the number of *citing sentences* and the number of *regular sentences* in the *citing sentence* predictor and the *regular sentence* predictor. Furthermore, the *reference-based* predictor and the combination predictor shared similar rank-1 features that were *num_ref_3years*. Second, an interesting fact here is that in the combination predictor, the rank-1, rank-2, and rank-3 features were filled by the rank-1 feature in the *reference-based* predictor, the *citing sentence* predictor, and the *regular sentence* predictor, respectively. This trend showed a consistent contribution of these rank-1 features in the regression tasks. Third, interestingly, the feature *citing_paper_dominant* was in the top 10 most important features in the *citing sentence* and *regular sentence* predictors, although the feature's distribution in the dataset is minimal. This trend corresponds with the phenomenon that occurs in the classification experiments.

Table 4.15. The best performance of average review score prediction for each regression scenario.

Predictors	**Average review Score**	**RFR**		**GBR**		**SVR**		**XGBR**		**DTR**	
		n	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*
Citing Sentence Predictor	RMSE	1	1.45	2	1.43	10	1.41	4	1.42	1	1.45
	MAE	1	1.17	4	1.14	6	1.12	4	1.14	1	1.17
Regular Sentence Predictor	RMSE	18	1.45	20	1.41	20	1.43	17	1.42	1	1.47
	MAE	20	1.16	20	1.14	20	1.15	20	1.14	1	1.19
Reference-based Predictor	RMSE	1	1.41	14	1.40	16	1.40	20	1.39	1	**1.41**
	MAE	1	1.14	16	1.11	23	1.12	23	1.11	1	**1.14**
Combination Predictor	RMSE	64	**1.34**	59	**1.34**	4	**1.37**	59	**1.34**	1	**1.41**
	MAE	64	**1.07**	59	**1.08**	59	**1.09**	63	**1.07**	1	**1.14**

** The bold values indicate the lowest result achieved by each algorithm.*

Table 4.16. The best performance of individual review score prediction for each regression scenario.

Predictors	**Individual review Score**	**RFR**		**GBR**		**SVR**		**XGBR**		**DTR**	
		n	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*
Citing Sentence Predictor	RMSE 1	15	1.82	10	1.80	11	1.81	10	1.80	1	1.84
	RMSE 2	19	1.91	5	1.87	6	1.90	10	1.86	1	1.92
	RMSE 3	20	2.05	3	2.04	20	2.09	4	2.04	1	2.06
	MAE 1	9	1.49	11	1.47	11	1.41	7	1.47	1	1.51
	MAE 2	20	1.58	11	1.54	10	1.49	10	1.54	1	1.59
	MAE 3	20	1.68	14	1.67	4	1.63	4	1.67	1	1.70
Regular Sentence Predictor	RMSE 1	17	1.83	2	1.81	19	1.81	1	1.81	1	1.86
	RMSE 2	18	1.90	7	1.88	1	1.93	19	1.87	1	1.95
	RMSE 3	16	2.05	5	2.03	3	2.09	6	2.02	1	2.13
	MAE 1	17	1.51	18	1.50	20	1.44	16	1.50	1	1.54
	MAE 2	8	1.57	8	1.54	30	1.52	8	1.54	1	1.61
	MAE 3	10	1.67	6	1.65	18	1.63	6	1.65	1	1.73
Reference-based Predictor	RMSE 1	1	1.79	17	1.76	17	1.78	2	1.77	1	1.79
	RMSE 2	1	1.89	20	1.87	19	1.90	19	1.87	1	1.89
	RMSE 3	1	2.02	17	1.99	24	2.05	18	1.99	1	2.02
	MAE 1	1	1.48	17	1.48	16	1.43	13	1.48	1	**1.49**
	MAE 2	21	1.54	20	1.54	24	1.52	24	1.54	1	1.55
	MAE 3	1	1.64	23	1.61	23	1.60	7	1.62	1	1.64
Combination All Predictor	RMSE 1	53	**1.71**	6	**1.72**	3	**1.73**	5	**1.73**	1	**1.79**
	RMSE 2	44	1.82	56	1.84	3	1.87	36	1.84	1	1.89
	RMSE 3	56	1.98	48	1.98	51	2.03	53	1.98	1	2.02
	MAE 1	53	**1.41**	26	**1.42**	4	**1.38**	22	**1.42**	1	**1.49**
	MAE 2	57	1.48	15	1.51	8	1.47	18	1.51	1	1.55
	MAE 3	56	1.61	62	1.60	6	1.58	60	1.60	1	1.64

** The bold values indicate the lowest result achieved by each algorithm.*

Table 4.17. The top 10 most influential features to achieve the best performances in both regression tasks.

rank	**#1 - citing sentence predictor**	**#2 - regular sentence predictor**	**#3 - reference-based predictor**	**#4 - combination predictor**
1st	number of citing sentence	number of regular sentences	num_ref_3years	#3 - num_ref_3years
2nd	citing_paper_use	citing paper corroboration	num_ref	#1 - number of citing sentence
3rd	other	citing paper use	iclr	#2 - number of regular sentences
4th	compare	citing paper dominant	neurips	#3 - num_ref
5th	citing paper corroboration	compare	icml	#2 - citing paper corroboration
6th	citing paper based on	other	arxiv	#1 - citing paper use
7th	citing paper dominant	contrast	neuralcom	#2 - citing paper use
8th	contrast	judgment	emnlp	#2 - citing paper dominant
9th	citing paper extend	suggest	acl	#1 - other
10th	judgment	citing paper based on	aistats	#1 - compare

Furthermore, evaluating the impact of features to achieve the best performance when using the combination predictor shows that the features belonging to the *citation functions*-based predictors dominated the distribution. Specifically, the distributions of *citing sentence* predictor, *regular sentence* predictor, and *reference-based* predictor in the top 10 most important selected features are 4, 4, and 2, respectively. Therefore, as previously mentioned in the classification tasks, the *reference-based* predictor contributes less to achieve the best performances when using a combination predictor.

This thesis compares the best results of regression tasks in this thesis with that of existing studies. Note that the comparison cannot be performed on all previous studies since most focused on predicting the aspect review scores (based on review comments) rather than the final review score. Therefore, the comparison can only be performed with the regression results from (Ribeiro et al., 2021) developed based on review comments that achieved the best RMSE and MAE of 1.28 and 1.05, respectively, which are slightly higher than the performances reached in this thesis. However, the best performances obtained in this thesis (RMSE: 1.34, MAE: 1.07) are considered competitive since the regression method was developed based on the paper without review comments.

4.5. Chapter Summary

This thesis developed a method for predicting paper quality to reduce the review burden that depends only on features extracted from the paper. This method is intended to handle the drawbacks of most existing studies involving the review comments for making the prediction. The prediction method proposed in this thesis encompasses three tasks where two are classification tasks, and the other is a regression task. The classification tasks primarily predict the paper quality to judge whether the submitted manuscripts are good or poor; however, the task of predicting the final review decision of accepted or rejected is also included for

comparison purposes. Conversely, the regression task can predict the average and individual review scores.

Furthermore, the experiments on the classification tasks demonstrate remarkable findings. First, predicting the paper quality based on the *good-poor* task is more effective than the *accepted-rejected* task. This was proved by error analysis results and supported by the achieved performances and the effectiveness, showing that the difference between TP-vs-TN and FP-vs-FN are separated in the *good-poor* task, although unclear in the *accepted-rejected* task. Second, the *citing sentences* predictor obtained a satisfactory performance by a recall of 0.99 in the *good-poor* task. Therefore, this result proves the hypothesis of this thesis concerning the crucial role of *citation functions* in the manuscript.

Regarding the regression experiment on the average and individual review scores, the combination predictor demonstrated its superiority over other predictors. However, *citing sentence* predictors showed a competitive performance using fewer classification features. These results increase the confidence level for making predictions by relying only on the paper when predicting the review scores.

Therefore, several points must be improved for further developments exist. First, it is worth applying the method in this thesis to other domains, e.g., broader CS and medicine, among others. Second, this thesis intends to explore more about using *citation functions* to predict the review aspect score (clarity, originality, impact, etc.) and the review score, which the assigned reviewers determine. Therefore, the developed prediction models to be one step closer to incorporating TAPR into the entire peer-review process.

Chapter 5 – Dataset of Citation Functions for COVID-19 Domain

This chapter explains the development of a new dataset of *citation functions* for COVID-19 academic papers. This dataset is proposed to address the necessity of comprehensive label of *citation functions* in the COVID-19 academic papers, which were not accommodated by previous works. This chapter contains several parts, i.e., existing works in this research area, dataset development, and dataset evaluation.

5.1. Existing Dataset of Citation Functions in COVID-19 Domain

There is a continuous development in designing labels for Rhetorical Structures (RS) and building datasets in the medical domain. Existing works have designed RS and developed the dataset (Alliheedi et al., 2019; Dayrell et al., 2012; Dernoncourt & Lee, 2017; Green, 2015; Jia, 2018; Kim et al., 2011; Liakata, 2010; Shatkay et al., 2008; Wilbur et al., 2006). However, several issues appear in these works. The first issue is that not all these RS were developed based on full text papers; several works built the RS using only papers' abstracts. The second issue is that most of the RS were not specifically designed for *citing sentences*. Since the existing RS covers *both citing sentences* and non-*citing sentences*, the number of labels is considered small, which causes several potential missing *citation functions* being accommodated—the last issue. Moreover, due to the COVID-19 pandemic, the number of published papers covering this topic has significantly increased. Existing RS is not designed specifically for this purpose, and this has become an additional issue. Considering this, this thesis aims to develop a new dataset of *citation functions* that contains more detailed labels, covers full text papers, and is specific for the COVID-19 domain.

Designing new labels of *citation functions* and building a new dataset is challenging. This is because it needs to provide large, labelled training data, which is time-consuming, expensive, and requires much human effort. To obtain the labeled instance with less effort, this thesis uses the labels of *citation functions* that have been built based on Computer Science (CS) papers. To realize this, this thesis uses the best model to classify the *coarse* labels and *fine-grained* labels of *citing sentences* in the CS domain. The model is then applied to categorize *citing sentences* on COVID-19-related papers obtained from the COVID-19 Open Research Dataset (CORD-19) (Lu Wang et al., 2020).

5.2. Dataset Development

This section consists of two parts. The first of which concerns the obtainment of data sources of COVID-19-related papers. The second part builds the dataset of *citation functions* on COVID-19 domains.

5.2.1. COVID-19-related Papers

This thesis uses a collection of papers from the COVID-19 Open Research Dataset (CORD-19) (Lu Wang et al., 2020). Initially, this dataset provided 28k papers. The present number of papers has significantly increased during continuous development. The CORD-192 collected papers from several sources (e.g., PubMed Central (PMC), PubMed, and the World Health Organization's COVID-19 Database). Moreover, it contains a collection from preprint servers such as bioRxiv, medRxiv, and arXiv. This thesis uses the latest version of the dataset (version: 2021-12-20) from JSON parsed from the full text of 314,391 (PDF) and 243,652 (PMC) papers. The distribution of CORD-19 is shown in Figure 5.1.

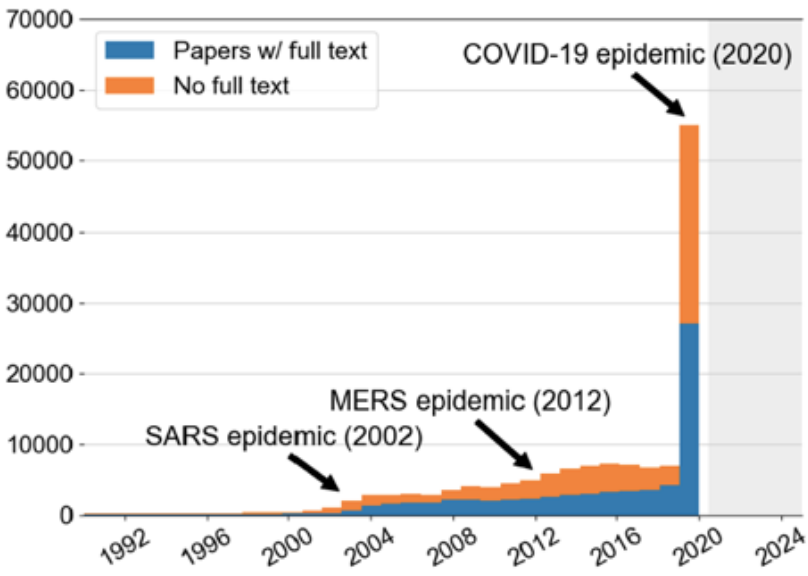


Figure 5.1. Paper distribution in the CORD-19.

** The x axis depicts year, and the y axis depicts the number of papers. This figure is taken from* (Lu Wang et al., 2020).

5.2.2. Dataset Development in COVID-19 Domains

The proposed dataset of COVID-19 domains is built using an automatic approach by following several steps. The first step is preparing the source of the papers. In this step, this thesis does a simple data analysis to gather a deep understanding of the parsed JSON structures of CORD-19. Following this, the analysis is accompanied by *filtering* to select only *citing sentences*. The second step is classifying all extracted *citing sentences* using the best models obtained from the dataset of the CS domain. The last step is verifying the automatically labeled *citing sentences* by performing a random selection of 475 instances and checking the predicted labels manually.

5.3. Experiment Results

This section explains the results of the automatic classification used to build a dataset of *citation functions* of the COVID-19 domain. The results are divided into several parts: brief information about classification models developed using the CS domain, the automatic classification of *citing sentences* of the COVID-19 domain, and the evaluation of classification through a manual label check. Note that the classification in the CS domain and the COVID-19 domain is done through two stages, namely the *filtering* stage and the *fine-grained* stage. While the *filtering* stage is used to classify the *citing sentences* into two categories, i.e., *Other (atr18)* and *No-Other (atr0-atr17)*, the *fine-grained* stage is applied to classify the *citing sentences* belonging to *No-Other* class into 18 *fine-grained* classes. Finally, the proportional distribution of labeled instances and a discussion of results are also presented.

5.3.1. A New Dataset of Citation Functions in COVID-19 Domain

The classification experiment is conducted on 99.6k instances generated from 10.1k parsed paper files (JSON format). The automatic classification begins with the extraction of all the sentences in the JSON files. Next, all extracted sentences are filtered to keep only *citing sentences*. Similar to the dataset on the CS domain, the classification is then applied by following two classification stages, namely the *filtering* stage and the *fine-grained* stage. To measure the accuracy of labeled instances, this thesis performs a manual label check on 25 random samples for each label, for a total of 475 samples (18 *fine-grained* labels + 1 other label).

After completing the manual label check, this thesis obtained accuracies 76.63% and 70.20% for *coarse*-grained labels and *fine-grained* labels, respectively. The accuracy of *coarse*-grained labels is easily obtained by summing the proportion of correctly and wrongly *fine-grained* labels. Since each label in the *fine-grained* labels has the same number of instances, it is easy to use the confusion matrix to compare each label's accuracy, as shown in Figure 5.2. The highest number of correctly predicted labels is achieved by the label technical, with 24 correct predictions and only a single incorrect prediction. In contrast, the label *cited_paper_dominant*

has the lowest number of correctly predicted labels with only nine correct and 16 incorrect predictions.

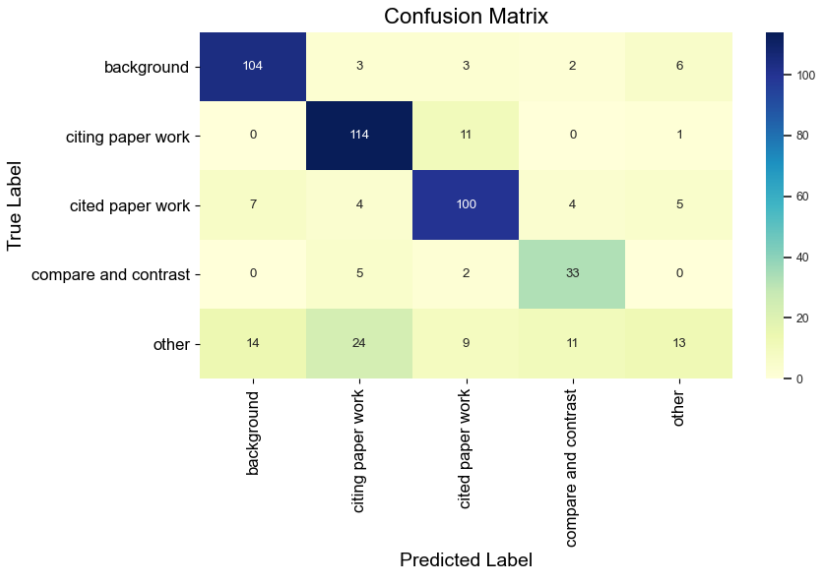


Figure 5.2.1. Coarse-grained labels.

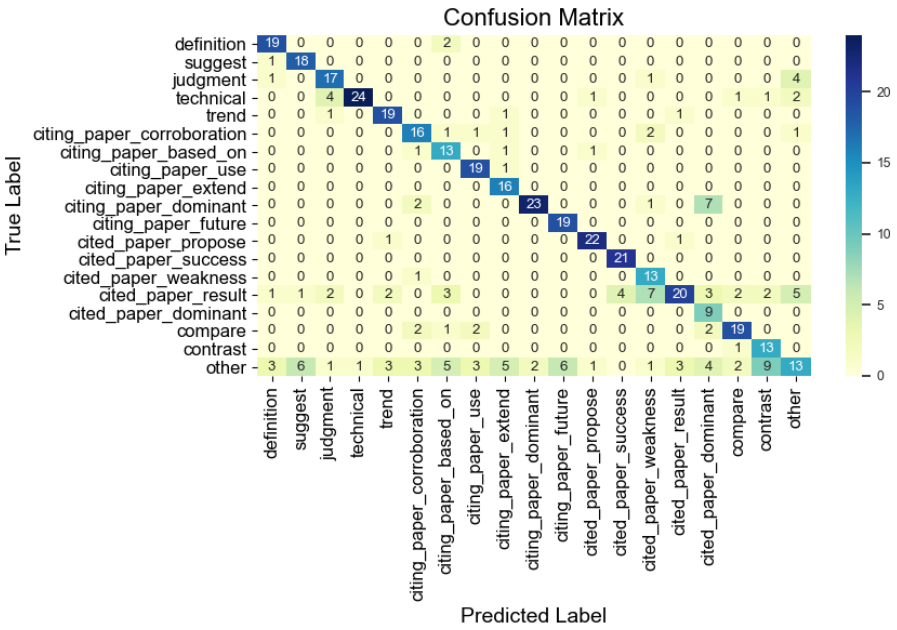


Figure 5.2.2. Fine-grained labels.

Figure 5.2. Confusion Matrix of manually label checking for (top) coarse-grained labels and (bottom) fine-grained labels.

Applying classification models built from CS papers to COVID-19 related papers results in two consequences. The first consequence is that there is a decrease of *fine-grained* label accuracy from 83.64% in CS domain to 70.2% in COVID-19 domain. The second consequence is that two *fine-grained* labels experienced a meaning shift: the label citing_paper_dominant and the label citing_paper_future. The definition of the label citing_paper_dominant changed from expressing the *citing paper*'s performance over *cited paper* to discussing the success of *citing paper*, with or without comparison. On the other hand, the definition of the label citing_paper_future changed from stating the future plan of the *citing paper* to a general recommendation without specifying whether it is done by *citing paper* or *cited paper*.

Table 5.1. The distribution comparison of automatically labeled instances in CS domain and COVID-19 domain.

Fine-grained Labels	**Number of Instances**		**Label Proportion**	
	CS Domain	**COVID-19 Domain**	**CS Domain**	**COVID-19 Domain**
definition	55,508	3,151	4.18%	3.77%
suggest	51,987	355	3.91%	0.42%
judgment	215,428	**37,885**	16.21%	**45.34%**
technical	85,374	5,557	6.42%	6.65%
trend	66,594	6,579	5.01%	7.87%
citing_paper_corroboration	113,488	2,571	8.54%	3.08%
citing_paper_based_on	55,878	531	4.20%	0.64%
citing_paper_use	115,215	1,114	8.67%	1.33%
citing_paper_extend	28,779	241	2.17%	0.29%
citing_paper_dominant	24,823	294	1.87%	0.35%
citing_paper_future	5,439	424	0.41%	0.51%
cited_paper_propose	**243,031**	5,442	**18.29%**	6.51%
cited_paper_success	34,505	2,128	2.60%	2.55%
cited_paper_weakness	15,054	1,072	1.13%	1.28%
cited_paper_result	154,394	15,063	11.62%	18.03%
cited_paper_dominant	**3,215**	**31**	**0.24%**	**0.04%**
compare	39,364	677	2.96%	0.81%
contrast	20,909	439	1.57%	0.53%
Total	**1,328,985**	**83,554**	**100%**	**100%**

** The comparison consists of two parts: (a) the number of instances on each label and (b) the proportion of instances on each label to the total instances in the dataset.*

5.3.2. The Distribution of Citation Functions in COVID-19 Domain

To give more analysis on the current COVID-19 dataset, Table 5.1 shows a comparison of the distribution datasets in the CS domain and COVID-19 domains. Note that the distribution in this table represents the number of automatically labeled *citing sentences* in the datasets. The current dataset in this thesis consists of 99,691 labeled instances, of which *No-Other* label has 83,554 instances and the *Other* label 16,137 instances. Since the labels of *citation functions* are designed for CS papers, it is worth determining whether the classification models are effective for domains related to COVID-19. Instead of using the number of instances to compare both datasets, this thesis uses the proportion of labels as indicators due to the datasets having different sizes.

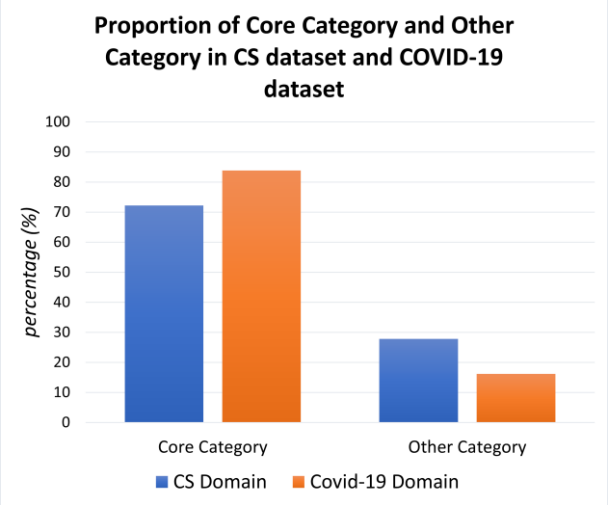


Figure 5.3. Proportion comparison between no-other and other labels.

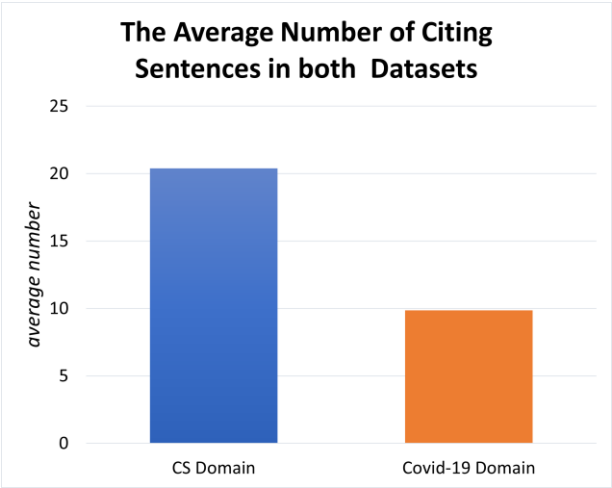


Figure 5.4. The average number of citing sentences in each paper.

First, the comparison is done on the *filtering* stage to show the percentage of *No-Other* vs *Other* labels as depicted in Figure 5.3. In this figure, it is seen that both domains share the same trend in that the proportion of *No-Other* label much higher than Other label. Surprisingly, the label judgment in the COVID-19 domain has a proportion of almost half at 45.34%. In second place, the label *cited_paper_result* has 18.03% of proportion. The rest of the labels constitute less than 10% of the proportion. Furthermore, there are eight labels that have only under 1% of the proportion, with the lowest proportion obtained by the label cited_paper_dominant with 0.04%, which is equivalent with 31 instances. The CS domain faces a similar situation in that this label has the lowest proportion at 0.24%. However, this proportion is not as severe as in the COVID-19 domain. In the dataset of CS domain, the distribution trend varies among labels, and no single label exceeds 20% of the proportion.

Another comparative indicator between both domains is the average number of *citing sentences* in each paper. Figure 5.4 demonstrates that the CS domain has a higher number of *citing sentences* than the COVID-19 domain. To be more specific, the dataset of CS domain consists of 1,840,815 *citing sentences* extracted from 90,278 papers, while the dataset of COVID-19 domain contains 99,691 *citing sentences* extracted from 10,102 papers.

5.4. Chapter Summary

The experiments conducted in this thesis reveal several notable findings. The first finding is a phenomenon of meaning shift in two *fine-grained* labels. This corroborates the assertion that even as this thesis achieves acceptable accuracies, there still exists an issue regarding the labels' compatibility between two domains. Next, the large proportion of label judgment (constituting almost half of dataset) indicates that *citation functions* in the COVID-19 papers are dominated with statements highlighting the importance, cruciality, usefulness, benefit, consideration, etc. of certain topics for making sensible argumentation. Conversely, the smallest proportion, represented by the label *cited_paper_dominant*, which is followed by several labels with proportions less than 1% (e.g., *compare*, *citing_paper_extend*, *contrast*, *citing_paper_dominant*, and *citing_paper_based_on*) indicates that discussing State of the Arts (SOTA) in the COVID-19 domain is less popular compared to the CS Domain. This trend is supported by the average number of *citing sentences* in the CS domain being higher than in the COVID-19 domain, which emphasizes the fact that discussing the SOTA needs more *citing sentences* and *cited papers*.

Chapter 6 – Summary and Conclusion

This thesis has developed a technology-assisted peer review based on *citation functions*, especially for predicting the paper quality and the final review decision. The prediction method is realized through several stages, i.e., building a new labeling scheme of *citation functions* based on Computer Science research papers, building dataset of *citation functions* using semi-automatic approach, and predicting the paper quality and final acceptance decision. In addition, this thesis has created a dataset of *citation functions* from COVID-19 academic papers.

The developed labeling scheme of *citation functions* consists of 5 *coarse* and 21 *fine-grained* labels. The proposed scheme was developed through top-down analysis, bottom-up analysis, and annotation experiments. The annotation experiments were validated using inter-annotator agreements using Kappa score which resulted in 0.85 (nearly perfect) for *coarse* labels and 0.71 (substantial agreement) for *fine-grained* labels. Besides the 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 dataset of *citation functions* in this thesis was developed using a semi-automatic approach. The final dataset consists of 1,840,815 labeled instances. During classification model development, BERT and SciBERT achieved higher accuracies than other methods. In addition, these two methods achieved promising results using active learning on less than half of the training data. SciBERT consistently outperformed BERT in the *fine-grained* stage in both active learning and non-active learning settings. However, BERT outperformed SciBERT in the *filtering* stage using active learning. Note that there is a consistent label distribution between the initial and final datasets.

This has developed a prediction system of paper quality which consists of two prediction tasks, namely *accepted-rejected* and *good-poor*. The experiments demonstrate notable findings. First, predicting the paper quality in terms of the *good-poor* task is more effective compared with the *accepted-rejected* task. Not only supported by the achieved performances, but the effectiveness is also proved by error analysis results that show the difference among TP-vs-TN and FP-vs-FN are clearly separated in the *good-poor* task but unclear in the *accepted-rejected* task. Second, the *citing sentences* predictor showed satisfying performance by 0.99 of recall in the *good-poor* task. This result proves the hypothesis related to the crucial role of citation in the manuscript.

This thesis has developed the dataset of *citation functions* using *citing sentences* extracted from COVID-19 related papers. Instead of designing new labels of *citation functions* from scratch and preparing training data, this thesis uses previously developed labels and applied the best ML models that have been built from the CS domain. The experiments show that the application of labels of the CS domain to the COVID-19 domain is promising. Furthermore, the evaluation for obtaining the automatic labeling accuracies uncovers several notable patterns such as label compatibility between two domains, the dominant citation roles on each domain, and the relation between a *citing paper* and the SOTA. For future work, this thesis intends to apply the labels and the models to all papers in the CORD-19 dataset.

Besides the benefit of using the proposed methods for TAPR, we identified several limitations. The proposed method promotes a specific style of paper writing in convincing the automatic prediction system rather than producing articles with sufficient quality. The next consequence is that since the citation functions based on Computer Science domain, the prediction method for paper quality only works for the same domain. Following this, the Feature Selection techniques for analyzing the top 10 most important features for predicting the paper quality are unable to provide the reason why these features were selected. These issues bring a new challenge for our future research in this domain.

References

Alliheedi, M., Mercer, R. E., & Cohen, R. (2019). Annotation of Rhetorical Moves in Biochemistry Articles. *Proceedings of the 6th Workshop on Argument Mining*, 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. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE)*, 131–168. https://doi.org/10.18293/SEKE2018-141

Bao, P., Hong, W., & Li, X. (2021). Predicting Paper Acceptance via Interpretable Decision Sets. *The Web Conference 2021 - Companion of the World Wide Web Conference (WWW 2021)*, 461–467. https://doi.org/10.1145/3442442.3451370

Basuki, S., & Tsuchiya, M. (2022). SDCF: semi-automatically structured dataset of citation functions. *Scientometrics*, *127*(8), 4569–4608. https://doi.org/10.1007/s11192-022-04471-x

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

Bharti, P. K., Ranjan, S., Ghosal, T., Agrawal, M., & Ekbal, A. (2021). PEERAssist : Leveraging on Paper-Review Interactions to Predict Peer Review Decisions. *International Conference on Asian Digital Libraries*, *1*, 421–435. https://doi.org/10.1007/978-3-030-91669-5

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

Breuning, M., Backstrom, J., Brannon, J., Gross, B. I., & Widmeier, M. (2015). Reviewer Fatigue? Why Scholars Decline to Review their Peers' Work. *PS - Political Science and Politics*, *48*(4), 595–600. https://doi.org/10.1017/S1049096515000827

Casey, A. J., Webber, B., & Glowacka, D. (2019). Can models of author intention support quality assessment of content? *Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019)*, 92–99. http://ceur-ws.org/Vol-2414/

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

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1). https://doi.org/10.1057/s41599-020-00703-8

Ciloglu, A., & Merdan, M. (2020). *Big Peer Review Challenge*. Humboldt-Universität.

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, *1*, 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

Dayrell, C., Candido, A., Lima, G., MacHado, D., Copestake, A., Feltrim, V. D., Tagnin, S., & Aluisio, S. (2012). Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 1604–1609.

Dernoncourt, F., & Lee, J. Y. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. *Proceedings Ofthe The 8th International Joint Conference on Natural Language Processing*, 308–313. http://arxiv.org/abs/1710.06071

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. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 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. *Proceedings of The 9th Linguistic Annotation Workshop*, 42–51. https://doi.org/10.3115/v1/W15-1605

Fox, C. W., Albert, A. Y. K., & Vines, T. H. (2017). Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review*, *2*(1), 1–6. https://doi.org/10.1186/s41073-017-0027-x

Fytas, P., Rizos, G., & Specia, L. (2021). What Makes a Scientific Paper be Accepted for Publication? *Proceedings of the First Workshop on Causal Inference and NLP*, 44–60. https://doi.org/10.18653/v1/2021.cinlp-1.4

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2019). DeepSentipeer: Harnessing sentiment in review texts to recommend peer review decisions. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1120–1130. https://doi.org/10.18653/v1/P19-1106

Ghosh, A., Pande, N., Goel, R., Mujumdar, R., & Sistla, S. S. (2020). *Prediction, Conference Paper Acceptance (Acceptometer)*. https://rohangoel.com/Acceptometer/;

Green, N. (2015). Identifying Argumentation Schemes in Genetics Research Articles. *Proceedings of the 2nd Workshop on Argumentation Mining*, 12–21. https://doi.org/10.3115/v1/w15-0502

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

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

Hu, P., Lipton, Z. C., Anandkumar, A., & Ramanan, D. (2019). ACTIVE LEARNING WITH PARTIAL FEEDBACK. *The International Conference on Learning Representations (ICLR)*, 1–14.

Jana, S. (2019). A history and development of peer-review process. *Annals of Library and Information Studies*, *66*(4), 152–162.

Jen, W., & Chen, M. (2018). *Predicting Conference Paper Acceptance* (pp. 1–7). Stanford University. https://cs229.stanford.edu/proj2018/report/117.pdf

Jia, M. (2018). Citation Function and Polarity Classification in Biomedical Papers. *The University of Western Ontario*. https://ir.lib.uwo.ca/etd/5367/

Johnson, R., Watkinson, A., & Mabe, M. (2018). *The STM Report - An overview of scientific and scholarly publishing* (Issue October). https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf

Joshi, D. J., Kulkarni, A., Pande, R., Kulkarni, I., Patil, S., & Saini, N. (2021). Conference Paper Acceptance Prediction: Using Machine Learning. *Machine Learning and Information Processing*, 143–152. https://doi.org/https://doi.org/10.1007/978-981-33-4859-2_14

Jubb, M. (2016). Peer review: The current landscape and future trends. *Learned Publishing*, *29*(1), 13–21. https://doi.org/10.1002/leap.1008

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

Kang, D., Ammar, W., Dalvi, B., Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1647–1661). https://doi.org/10.18653/v1/N18-1149

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, *12*(SUPPL. 2). https://doi.org/10.1186/1471-2105-12-S1-S5

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLoS ONE*, *5*(4), 2–6. https://doi.org/10.1371/journal.pone.0010072

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

Li, J., Sato, A., Shimura, K., & Fukumoto, F. (2020). Multi-task Peer-Review Score Prediction. *Proceedings of the First Workshop on Scholarly Document Processing*, 121–126. https://doi.org/10.18653/v1/2020.sdp-1.14

Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards fine-grained citation function classification. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, *September*, 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*, 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*, 233–250. https://doi.org/https://doi.org/10.1007/978-981-15-4771-3_10

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., … Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Maillette de Buy Wenniger, G., van Dongen, T., Aedmaa, E., Kruitbosch, H. T., Valentijn, E. A., & Schomaker, L. (2020). Structure-tags improve text classification for scholarly document quality prediction. *Proceedings of the First Workshop on Scholarly Document Processing*, 158–167. https://doi.org/10.18653/v1/2020.sdp-1.18

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/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. *Advances in Artificial Intelligence - 17th Conference of the Canadian Society for Computational Studies of Intelligence*, *3060*, 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. *Conference on Neural Information Processing Systems*.

Moher, D., Galipeau, J., Alam, S., Barbour, V., Bartolomeos, K., Baskin, P., Bell-Syer, S., Cobey, K. D., Chan, L., Clark, J., Deeks, J., Flanagin, A., Garner, P., Glenny, A. M., Groves, T., Gurusamy, K., Habibzadeh, F., Jewell-Thomas, S., Kelsall, D., … Zhaori, G. (2017). Core competencies for scientific editors of biomedical journals: Consensus statement. *BMC Medicine*, *15*(1). https://doi.org/10.1186/s12916-017-0927-0

Nazir, S., Asif, M., & Ahmad, S. (2020). Important Citation Identification by Exploiting the Optimal In-text Citation Frequency. *2020 International Conference on Engineering and Emerging Technologies, ICEET 2020*. https://doi.org/10.1109/ICEET48479.2020.9048224

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

Pierson, C. A. (2018). Peer review and journal quality. In *Journal of the American Association of Nurse Practitioners* (Vol. 30, Issue 1, pp. 1–2). Lippincott Williams and Wilkins. https://doi.org/10.1097/JXX.0000000000000018

Pradhan, T., Bhatia, C., Kumar, P., & Pal, S. (2021). A deep neural architecture based meta-review generation and final decision prediction of a scholarly article. *Neurocomputing*, *428*, 218–238. https://doi.org/10.1016/j.neucom.2020.11.004

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. *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*, 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). http://informationr.net/ir/21-4/paper733.html

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

Ribeiro, A. C., Sizo, A., Lopes Cardoso, H., & Reis, L. P. (2021). Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis. *EPIA 2021: Progress in Artificial Intelligence*, *12981 LNAI*, 766–777. https://doi.org/10.1007/978-3-030-86230-5_60

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

Rowland, F. (2002). The peer-review. *Learned Publishing*, *15*(4), 247–258. https://doi.org/https://doi.org/10.1087/095315102760319206

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

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: Randomised controlled trial. In *British Medical Journal* (Vol. 328, Issue 7441, pp. 673–675). BMJ Publishing Group. https://doi.org/10.1136/bmj.38023.700775.ae

Settles, B. (2010). Active Learning Literature Survey. In *Computer Sciences Technical Report 1648*. 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

Shatkay, H., Pan, F., Rzhetsky, A., & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, *24*(18), 2086–2093. https://doi.org/10.1093/bioinformatics/btn381

Skorikov, M., & Momen, S. (2020). Machine learning approach to predicting the acceptance of academic papers. *The 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 113–117. https://doi.org/10.1109/IAICT50021.2020.9172011

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.

Stappen, L., Rizos, G., Hasan, M., Hain, T., & Schuller, B. W. (2020). Uncertainty-aware machine support for paper reviewing on the interspeech 2019 submission corpus. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1808–1812. https://doi.org/10.21437/Interspeech.2020-2862

Su, X., Prasad, A., Sugiyama, K., & Kan, M. Y. (2019). Neural multi-task learning for citation function and provenance. *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2019-June*, 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*, *121*, 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Tennant, J. P. (2018). The state of the art in peer review. *FEMS Microbiology Letters*, *365*(19), 1–10. https://doi.org/10.1093/femsle/fny204

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

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

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

Tong, Z., Huan, Y., Lei, S., Jing, W., & Daojia, X. (2021). Application and classification of artificial intelligence-assisted academic peer review. *Chinese Journal of Scientific and Technical Periodals*, *32*(1), 65–74. https://doi.org/10.11946/cjstp.201911220799

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*, 1041–4347. https://doi.org/DOI 10.1109/TKDE.2019.2913376

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

Vincent-Lamarre, P., & Larivière, V. (2021). Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. *Quantitative Science Studies*, *2*(2), 662–677. https://doi.org/10.1162/qss_a_00125

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, K., & Wan, X. (2018). Sentiment analysis of peer review texts for scholarly papers. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 175–184. https://doi.org/10.1145/3209978.3210056

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

Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., & Rajani, N. F. (2020). ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In B. Davis, Y. Graham;, J. Kelleher;, & Y. Sripada; (Eds.), *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 384–397). Association for Computational Linguistics. https://aclanthology.org/2020.inlg-1.44

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines and corpus construction. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-356

Yuan, W., Liu, P., & Neubig, G. (2022). Can We Automate Scientific Reviewing? *Journal of Artificial Intelligence Research*, *75*, 171–212. https://doi.org/https://doi.org/10.1613/jair.1.12862

Zhao, H., Luo, Z., Feng, C., & Ye, Y. (2019). A Context-based Framework for Resource Citation Classification in Scientific Literatures. *SIGIR'19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1041–1044. https://doi.org/10.1145/3331184.3331348

Publication List

Basuki, S., Tsuchiya, M. Automatic Approach for Building Dataset of Citation Functions for COVID-19 Academic Papers. In Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) (2022).

Basuki, S., Tsuchiya, M. SDCF: semi-automatically structured dataset of citation functions. Scientometrics 127, 4569–4608 (2022).

Basuki, S., Tsuchiya, M. The Quality Assist: A Technology-Assisted Peer Review based on Citation Functions to Predict the Paper Quality. IEEE Access (2022).