IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 October 2022, accepted 21 November 2022, date of publication 1 December 2022,
date of current version 8 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225871

RESEARCH ARTICLE

The Quality Assist: A Technology-Assisted Peer Review Based on Citation Functions to Predict the Paper Quality

SETIO BASUKI AND MASATOSHI TSUCHIYA
Department of Computer Science and Engineering, Toyohashi University of Technology (TUT), Toyohashi, Aichi 441-8580, Japan
Corresponding author: Setio Basuki (setio@is.cs.tut.ac.jp)

This research was supported by Toyohashi University of Technology (TUT) – Japan; in part by Amano Institute of Technology Scholarship; in part by JSPS KAKENHI Grant Number JP22K12167.

ABSTRACT This study aims to develop a prediction model for paper quality assessment to support technology-assisted peer review. The prediction technique is intended to reduce the review burden, which is becoming a critical issue in today's paper submission process. However, most existing works on this topic were built by involving the reviewers' comments, which is considered unfair and inapplicable for reducing the review burden. Therefore, our prediction method relies only on features extracted from the paper to address this issue. The method covers three tasks as follows: two are classification tasks and one is a regression task. The classification tasks predict the final review decision (accepted-rejected) and estimate the paper quality (good-poor), while a regression task predicts the review scores. Additionally, the classification and regression tasks are implemented using three main features i.e., citing sentence features developed based on the labeling scheme of citation functions, regular sentence features created by applying the label of citation functions to non-citation text, and reference-based features constructed by identifying the source of citations. Furthermore, the classification experiments on the dataset obtained from the International Conference on Learning Representations 2017–2020 showed that our methods are more effective in the good-poor task than the accepted-rejected task by demonstrating the best accuracy of 0.75 and 0.73, respectively. Moreover, we also reached a satisfactory recall of 0.99 using only the citing sentence features to obtain as many good papers as possible in the good-poor task. Our regression experiments indicate that the best result in predicting the average review score is higher than the individual review score by showing Root Mean Square Error (RMSE) of 1.34 and 1.71, respectively.

INDEX TERMS Citation function, final review decision, paper quality, review score, technology-assisted peer review.

I. INTRODUCTION

Peer review aims to ensure the quality of scientific works. It is used not only in journal publishing but also in conference submissions, grant proposal evaluations, and academic monograph submissions [1]. However, completing all stages of the peer review is time-consuming and requires extensive human effort, from accepting the manuscript to the final review decision. Peer review can be challenging in the journal submission process due to the massively published research papers. The STM report 2018 [2] states 33,100 peer-reviewed English-language journals and 9,400 non-English-language journals collectively publish more than 3 million articles annually. Another study reported that the yearly review of the previously rejected manuscripts reaches 15 million hours [3]. Moreover, EasyChair, a web application for conference management systems, has managed around 100,000 conference events since 2002.[1] The situation worsens due to

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

[1] https://easychair.org/

the uneven geographical distribution of the review experts, thereby putting the peer-review process into an over-burdened system [4].

Another limitation of the peer-review process is that because it is based only on human expertise, it will unavoidably tend to be biased and subjective due to several factors, such as expert academic background, experience, emotion, and health [5]. Other challenges well identified by [6] include inadequate training on how to perform the peer review or response to the review [7], the relationship between the journal and peer-review quality [8], and standard core competencies for editors [9]. Additionally, Jana [10] explained additional limitations of the traditional peer-review system, such as expensive and publication delay, harsh comments due to reviewers' anonymity, author-recommended reviewers, and irresponsible reviewers to complete the review process. Therefore, this situation poses the opportunity for proposing a technology-assisted peer review (TAPR), an automated screening method, to reduce the massive burden of the peer review process.

The development of TAPR to reduce the review burden has gained much attention. The TAPR has addressed three principal tasks in existing works: predicting the paper quality, final review decisions, and review scores. The exiting TAPR was developed using various purposes, ranging from predicting three-classes outcomes accepted, borderline, and rejected [11] to suggesting two-labels outcome accepted and rejected as majority-targeted classes as in [12]. However, existing works encounter two main drawbacks. **The first drawback** is that due to inconsistency among review results, review scores, and final decisions, as stated by [13], directly predicting the final review decision leads to a bias in determining the paper's quality. For example, if reviewers agreed not to reject the manuscripts, the editor rejected 20%, or if the reviewers agreed to reject the manuscripts, the editor rejected 80%. Therefore, to resolve this issue, this paper proposes two prediction tasks: the paper quality prediction to determine whether the manuscripts are good or poor, which is more reasonable and review score prediction to estimate the review scores. However, for comparison, we predict that the final reviewer's decision based on the submitted manuscripts will be accepted or rejected. **The second drawback** is that most existing studies employed the review comments as prediction features. However, this approach is considered unfair and inapplicable when the main aim is to reduce the review burden. Therefore, the technique to reduce the human cost of the peer-review process should not depend on the features, including review comments that require human work.

This study develops a prediction method to address two classification tasks and a regression task for assessing paper quality; these tasks do not depend on the review comments. While the classification tasks predict the final review decision (*accepted-rejected*) and paper quality (*good-poor*), the regression task forecasts the average and the individual review scores. Additionally, the prediction tasks are accomplished using predictors with several prediction features. Here, we use *citation functions*, which represent why the author of the research paper cited previous works as the main predictor. This choice is motivated because the *citation functions* can represent the paper's quality [14] [15], show the proposed research position in numerous literature [16], indicate the novelty of the proposed research [17], understand the broad view of the given research topics of the paper [18], examine the map of science [19]. Additionally, the use of citation functions-based features brings another advantage to explore rarely-touched field of citation functions-based recommendation system, especially when preparing the research manuscript. Hence, the significant role of citation functions in estimating the paper's quality is worth discussing.

The prediction method for the tasks proposed in this paper can be summarized as follows: the main predictor is *citation functions* obtained by categorizing the *citing sentences* (a sentence containing citation marks). Notably, the *citation functions* applied in this paper were developed in our previous research [20]. Since the author's intention during manuscript writing cannot be accommodated using only *citing sentences*, this paper proposes an additional predictor called *regular sentence* predictor involving the non-citation sentences. Following this, another predictor to be implemented here is the *reference-based* predictor, which represents the references cited in the manuscript. Finally, we intend to merge the mentioned predictors into a combination predictor to investigate the impact of prediction features when combined. The prediction model is created using several Machine Learning (ML) and Feature Selection (FS) methods. Therefore, to evaluate the prediction performances, this study uses a dataset from the International Conference on Learning Representations (ICLR) 2017–2020, well parsed by [21]

At the end of this paper, several contributions will be explained:

- This paper proposed a method to predict the final review decision, paper quality, and review scores comprising four predictors as follows: *citing sentence*, *regular sentence*, *reference-based*, and combination. Our prediction method is independent of review comments as features but depends only on the paper.
- This paper demonstrated the accuracy of 0.67 and 0.72 in accepted-rejected and good-poor tasks, respectively, to evaluate the impact of *citation functions* in the classification tasks. However, the best accuracies were achieved through a combination predictor by 0.73 and 0.75 in accepted-rejected and good-poor tasks, respectively.
- Compromising with lower accuracy of 0.72 in the good-poor task, the satisfying recall of 0.99 was achieved using only *citing sentence* predictor.
- Analyzing the top 10 most important features in the combination predictor of classification tasks poses the fact that a feature called *citing_paper_dominant,* which represents a paper that outperforms previous works' performance, is considered significant to the prediction

results; however, this feature has few instances of distributions in the dataset.
- Regarding average review score prediction, the best results were represented by RMSE and Mean Absolute Error (MAE), achieving 1.34 and 1.07. Conversely, in the individual review score prediction, the best RMSE and MAE were attained when predicting the individual score 1 by 1.71 and 1.38.
- When obtaining the best performance, the *citation functions*-based predictors (*citing sentence* and *regular sentence* predictor) are more impactful than the *reference-based* predictor.
- Finally, our prediction method is more effective in predicting the paper quality than the final review decision in the classification tasks. However, in the regression task, our method better estimates the average review score than the individual one.

This paper is organized as follows: Section II presents a brief review of related works predicting the paper's final review decision, paper quality, and review score. Section III introduces our proposed method to handle three prediction tasks, i.e., *accepted-rejected*, *good-poor,* and review scores. Next, Section IV describes how the prediction features are constructed. In Section V, we report the experimental results of the prediction tasks. Finally, Section VI encompasses the conclusion and future research plan.

II. RELATED WORK

This section presents existing works on three research focuses as follows: (a) existing TAPR platforms were developed by publishers and technology vendors, (b) paper acceptance prediction covering two subtasks: classification tasks, comprising the final review decision and the paper quality prediction, and regression tasks for predicting review scores, and (c) limitation of existing predictions methods. Finally, we highlight the limitations of how existing prediction methods were developed and illustrate the contribution of this paper to this research area.

A. EXISTING TAPR PLARFORMS

The TAPR tools have been developed by both publishers and technology vendors for different purposes. For example, Frontiers 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 follow 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 Institute of Health (NIH), Materials Design Analysis Reporting (MDAR), Animal Research: Reporting of In Vivo Experiments (ARRIVE), etc. and the provides scores for every submission. Following this, Scholastica[5] optimize the peer review through integrating the peer review itself with the production and journal hosting software. Elsevier has released editorial tool called EVISE[6] for several tasks including plagiarism detection and reviewer matching. All these developed tools proofs that the peer review system needs to be intervened by technologies to solve the issues of review burden.

B. PAPER ACCEPTANCE PREDICTION

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 [22] is the most cited work, which [22] 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 [15] which used the 94 Related Work section of the ACL dataset, and [23] 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 the 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 [11] which used two and three labels for *accepted-rejected* and *accepted-borderline-rejected*, respectively, and in [15] 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 [22]. Additionally, two existing studies proposed the final review scores as in [23] and [24].

C. LIMITATTION OF EXISTING PREDICTION METHODS

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

[2]https://blog.frontiersin.org/tag/aira/
[3]https://discovery.researcher.life/publisher
[4]https://sciscore.com/
[5]https://scholasticahq.com/features/
[6]https://www.elsevier.com/connect/reviewers-update/rolling-out-our-new-editorial-system-evise

TABLE 1. Existing studies on final review decision and paper quality.

Paper	Title	Dataset for Prediction	Feature	Focused Task
[22]	A Dataset of Peer Reviews (PeerRead): Collection, Insights, and NLP Applications	ICLR 2017, ArXiv	hand-engineered coarse and lexical features.	Accepted-rejected
[11]	Sentiment analysis of peer-review texts for scholarly papers	ICLR 2017–2018	review comments.	Accepted-rejected (2 classes); and accepted-borderline-rejected (3 classes)
[25]	Predicting Conference Paper Acceptance	Using ICLR 2017	pre-defined features.	Accepted-rejected
[15]	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)
[24]	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
[26]	Conference Paper Acceptance Prediction (Acceptometer)	ICLR 2017	the paper, review comment.	Accepted-rejected
[27]	Machine learning approach to predicting the acceptance of academic papers	NIPS, ICLR, ACL, CoNLL, ArXiv (treated as single dataset)	pre-defined features.	Accepted-rejected
[28]	Structure-tags improve text classification for scholarly document quality prediction	ArXiv	tag structure of the paper.	Accepted-rejected
[29]	Big Peer-Review Challenge	NIPS, ICLR, ACL, CoNLL, ArXiv (treated as single dataset)	the paper, review comment.	accepted-rejected
[30]	Conference Paper Acceptance Prediction: Using Machine Learning	ICLR 2017	pre-defined coarse and lexical features.	Accepted-rejected
[31]	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
[32]	Predicting Paper Acceptance via Interpretable Decision Sets	ICLR 2017, ArXiv	pre-defined features.	Accepted-rejected
[23]	Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis	AI Conference 2013 & 2019, and Robotics 2015 and 2019	review comment.	Accepted-rejected
[33]	PEERAssist: Leveraging on Paper-Review Interactions to Predict Peer-Review Decisions	ICLR 2017–2020	the paper, review comment, review sentiment.	Accepted-rejected
[12]	What Makes a Scientific Paper be Accepted for Publication?	ICLR 2017	review comment, meta review.	Accepted-rejected
[34]	A deep neural architecture based meta-review generation and final decision prediction of a scholarly article	ICLR 2017–2019	review comment.	Accepted-rejected

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 paper 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. We propose 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, we 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 [11] 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, we prefer 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. [15] 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 1 and 2 Show the details of the existing studies.

III. PREDICTION METHOD

This method briefly describes the stages used to build the prediction method proposed in this paper, as shown in Figure 1. The prediction method follows several stages: In the **first stage**, we discuss the research papers' data source, which is a paper acceptance dataset. **The second stage** explains three

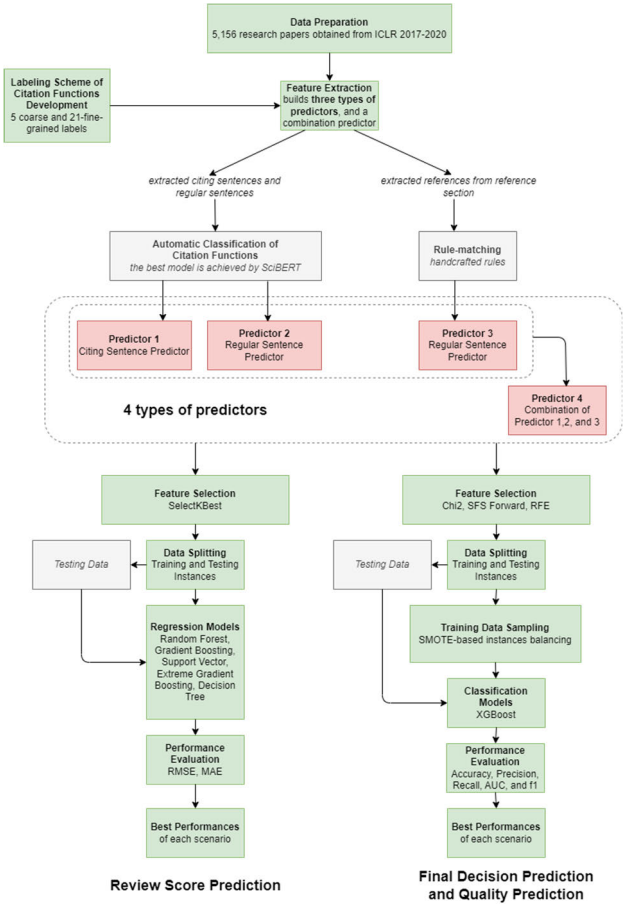


FIGURE 1. The general architecture of the proposed method for both classification tasks and regression task.

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.

Therefore, we define several terminologies used in the entire paper for consistency. These terms include *citing paper as* an author's work; *citing paper* as previous work cited by the *citing paper; citing sentence* as a sentence containing *citation marks; and a regular sentence* that does not contain citation marks. Therefore, we introduce the term *predictor* as several classification features. This section explains the three types of predictors, including *citing sentences*, *regular sentences*, and *reference-based* predictors. The other parts of the proposed method will be explained in the next section.

A. 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 explain why the author of the research papers cited previous works. Therefore, we use the labeling scheme of *citation functions* developed in our previous study [20] comprising 5 *coarse* and 21 *fine-grained* labels. The scheme of *citation function* was developed using a research paper dataset from [38], containing 90,278 parsed papers from arXiv Computer Science (CS) from January 1993 to December 31, 2017. Furthermore, we define *coarse* labels 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 we include one more feature 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, we denote the features as *c0* to *c19* for encoding purposes, as shown in Table 3.

B. REGULAR SENTENCE PREDICTOR

The *regular sentence* predictor is the first additional predictor proposed in this paper. 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 we categorize all *regular sentences* extracted from each paper of the ICLR dataset using ML when classifying the *citing sentences*. Therefore, this predictor will have the same labels as the *citing sentence* predictor, and we denote the labels starting from *r0* to *r19*.

C. REFERENCE-BASED PREDICTOR

The second additional predictor proposed in this paper 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 our 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 our dataset. Additionally, the review shows that the papers are frequently cited in preprint repositories and references published within 3 years. Therefore, we encode the labels from *ref0* to *ref23* and all the labels as prediction features. Table 4 presents detailed features of this predictor.

D. COMBINATION PREDICTOR

Here, we include 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. We denote the

TABLE 2. Existing works on review score prediction.

Authors	Paper Title	Category	Dataset	Features
[22]	A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications	Aspect review score prediction	ACL 2017; ICLR 2017	Review comments
[24]	DeepSentiPeer: Harnessing Sentiment in Review Texts To Recommend Peer-Review Decisions Tirthankar	Final review score prediction	ICLR 2017; ACL 2017; CoNLL 2016	Review comments
[35]	Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus	Aspect review score prediction	Interspeech 2019	Review comments
[36]	ReviewRobot: Explainable Paper-Review Generation based on Knowledge Synthesis	Aspect review score prediction	ACL 2017	Review comments
[37]	Multi-task Peer-Review Score Prediction	Aspect review score prediction	ICLR 2017 and ACL 2017	Paper text
[23]	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)

features in this predictor as *comb0* to *comb63* for the encoding purpose.

IV. 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 paper. Secondly, this section discusses the creation of prediction features and their distribution. Lastly, this section describes how the experiment scenarios are planned and executed.

A. THE DATASET OF PAPER ACCEPTANCE

This paper applies the dataset from [21], 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 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. We decided the papers had $4<$average review scores<5 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, we do not consider whether the accepted paper will be presented as an oral, poster, or workshop. The assumption in using the review score as the quality indicator is that the reviewers have already considered several review aspects such as originality, novelty, clarity, impact, etc. as a common guidance when doing the review. This paper selected 5,156 papers out of 5,192 papers from the dataset. This difference occurs because we 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. We also identified 3,575 and 1,581 *good* and *poor* papers, respectively, within the same dataset. Table 5 shows the detailed dataset distribution.

B. BUILDING THE CLASSIFICATION FEATURES

The classification features are created by gathering each feature (label) of all predictors in the paper. Therefore, we extract 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 our developed ML model based on SciBERT [39] obtained from our previous study [20]. Accordingly, our SciBERT model achieved an accuracy of 0.83, followed by an f1 score of 0.84. We 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[7] python package. Conversely, for the *reference-based* predictor, we employed the keyword matching approach to estimate each label in all papers. Therefore, to create the combination predictor, we simply merge 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*.

C. BUILDING THE REGRESSION FEATURES

The review score prediction applies similar features as that in the classification tasks. The difference is that the review score

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

TABLE 3. The coarse and fine-grained labels of citation function 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.

TABLE 3. *(Continued.)* The coarse and fine-grained labels of citation function as a list of features in the citing sentence predictor.

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>.

prediction is considered a regression problem comprising two tasks, i.e., average and individual review score predictions. The average review score is 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, we treat the average and the individual review score predictions as single-and multi-output regressions, respectively. Therefore, l both regression tasks will follow similar experiment settings.

D. 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 our method.

Here, we discuss the instance distribution of all predictors. Table 6 shows the yearly distribution. Figure 2 depicts the distribution of entire years. In Figure 2.1 and Figure 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 using labels in the *regular sentence* predictor adopted from the *citing sentence*. In Figure 2.3, the spread of labels in the *reference-based* predictor is dominated by the number of references for the last 3 years *(NUM-REF2YEARS)*, followed by preprint source *(arXiv)*, *ICLR, NeurIPS,* and *ICML*. Furthermore, the other labels in this predictor possess relatively equal distribution.

Fig. 3 demonstrates the comparison of the mean distribution of all predictors. Notably, the relatively equal distribution happens in the citing sentence and the reference-based predictors. Generally, the distribution of regular sentence predictors should be significantly higher than the other two predictors.

TABLE 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

TABLE 5. Distribution of paper collection used in this paper.

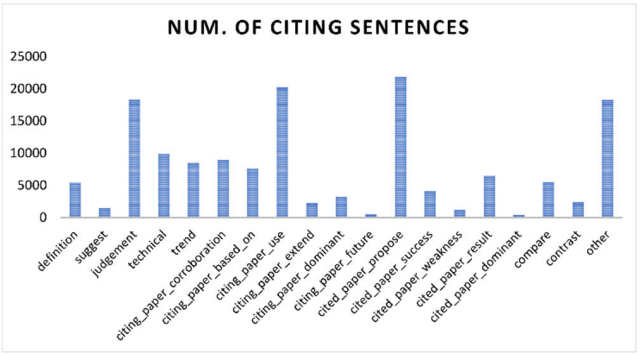
Year	Accepted Papers	Rejected Papers	Good Papers	Poor Papers	Total
2017	198	289	416	71	487
2018	336	571	769	138	907
2019	502	1,048	1,275	275	1,550
2020	686	1,526	1,115	1,097	2,212
Total	1,722	3,434	3,575	1,581	**5,156**

TABLE 6. Distribution of each predictor in the dataset.

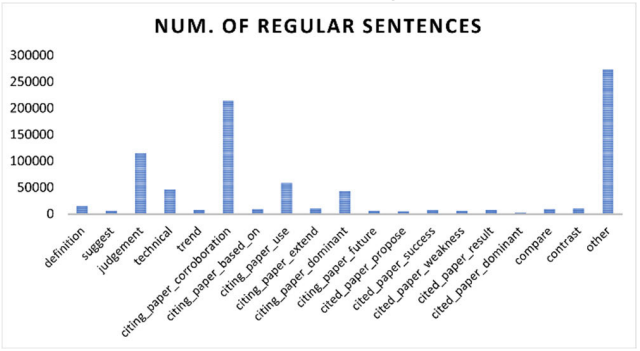
Paper Sources	Num. of Papers	Type of Classification Features	Num. of Instances
ICLR 2017	487	Citing Sentences	12,250
		Regular Sentences	72,940
		Reference-based	13,844
ICLR 2018	907	Citing Sentences	24,238
		Regular Sentences	141,741
		Reference-based	27,670
ICLR 2019	1,550	Citing Sentences	43,690
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

E. EXPERIMENT SCENARIO

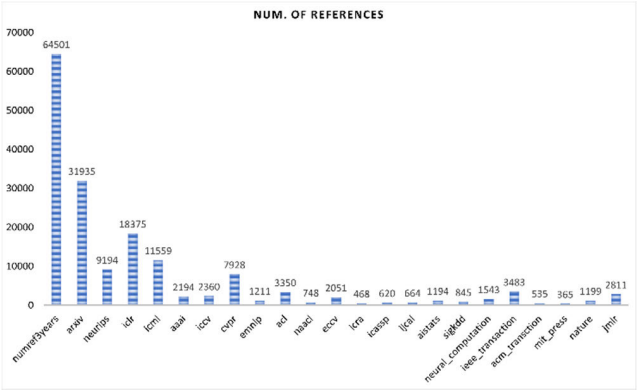
Here, the *accepted-rejected* and *good-poor* predictions are treated as classification issues. Both prediction tasks apply similar experimental settings as follows: we propose four experiment scenarios, with each scenario representing each type of predictor. Specifically, the experiment on the *citing sentence, regular sentence, reference-based, and combination* predictors adopt features *c0* to *c19*, *r0* to *r19*, *ref0* to *ref23*, and *comb0* to *comb63*, respectively. We apply XGBoost 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.[8] The FS



(a) Distribution of citing sentences.



(b) Distribution of regular sentences.



(c) Distribution of references.

FIGURE 2. The distribution of all classification features in ICLR from 2017 to 2020 is presented on each attribute. In Figure 2.3, we did not present the number of reference distributions (NUM_REF) because it is obtained by accumulating all other labels' distributions.

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

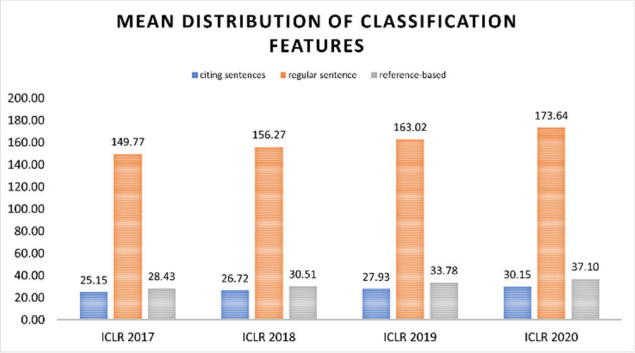


FIGURE 3. The means' distribution of the citing and regular sentences in the ICLR datasets. Here, the x and y axes represent the sentences' categories and means, respectively.

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, we evaluate the data balancing technique's impact on the classification performances using Synthetic Minority Over-sampling Technique (SMOTE)-based method.[9]

Conversely, this paper 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.

V. 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 our study and previous studies.

A. PERFORMANCE OF CLASSIFICATION TASKS

Tables 7 and 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 as Chi2 (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

[9]https://imbalanced-learn.org/stable/

TABLE 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

TABLE 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

Note: the bold values indicate the best of the best of all scenarios.

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.

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

TABLE 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

TABLE 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*.

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.

1) 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 9 and 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 11, and Table 12 shows the detailed selected features in the *accepted-rejected* and *good-poor* tasks, respectively.

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.

We report 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 2.1 and Figure 2.2).

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 2.3).

Another fascinating finding in our 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 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

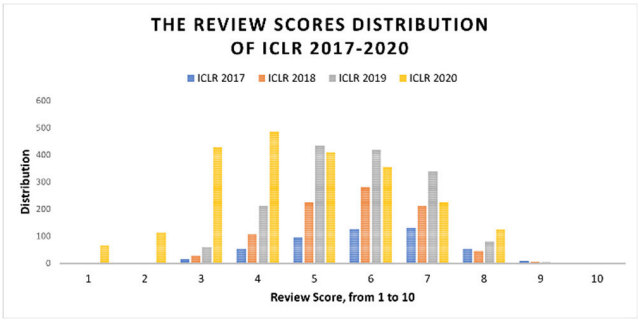


FIGURE 4. Distribution of review scores in ICLR from 2017 to 2020.

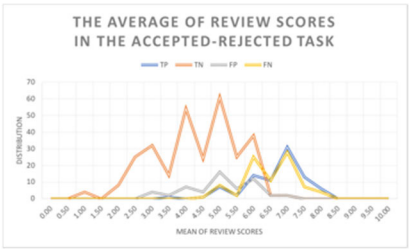
when using the imbalanced setting, the balanced setting for the same task or both imbalanced and balanced settings on good-poor task had accuracy reaching the same or even lower results compared with *citation functions*-based predictors. Altogether, the *reference-based* predictor still contributes to forming the combination predictor, although the *citation functions*-based predictors have more impact in obtaining the best results.

2) ANALYSIS TOWARD THE REAL REVIEW SCORES OF CLASSIFICATION EXPERIMENTS

It is worth discussing why our models were effective in the *good-poor* task rather than the *accepted-rejected* task. Accordingly, we depict the review scores of ICLR 2017–2020 in Figure 4 and the mean and variance of review scores of the best results in both classification tasks in Figure 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, our proposed classification features are more effective at categorizing whether the paper is good or poor rather than predicting its acceptance rate.

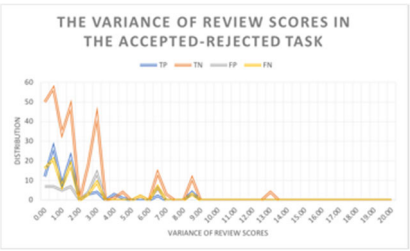
3) THE MEANING SHIFT OF REGULAR SENTENCE PREDICTOR

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



(a) The average review scores in accepted-rejected task

(b) The average review scores in good-poor task



(c) The variance of review score in accepted-rejected task

(d) The variance of review score in good-poor task

FIGURE 5. Distribution of Mean of review Scores and Variance of review Score of the best results in Accepted-Rejected and Good-Poor Tasks.

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 13 presents a detailed explanation of this phenomenon.

4) PERFORMANCE COMPARISON OF CLASSIFICATION EXPERIMENTS IN THIS PAPER WITH PREVIOUS WORKS

Here, the performance comparison cannot be conducted on the same dataset. This because there is no single standard of benchmark dataset which has final review decision and review scores as comprehensive as provided by ICLR. For example, there are works that use datasets only for prediction of final review decision based on arXiv using two classes: *accepted* vs *probably-rejected*. Since directly predicting the

TABLE 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 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 11 and Table 12, we use mark # 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.

final decision is problematic, we propose not only predicting the final decision but also predicting the paper quality and review scores. Therefore, the comparison in our paper is presented to show that the performances our method are competitive compared with previous works even though not using the reviewers' comments.

Generally, several existing works used accuracy as the only performance metric. Two studies employed alternative metrics, such as [31] using the f1, and [32] which employed the AUC. The other three studies employed more than one metric such as [27] which used accuracy, precision, recall, and f1, [28] which used accuracy and AUC, and [23] which used accuracy, recall, and f1. Here, we applied five metrics, i.e., accuracy, precision, recall, f1, and AUC (see Tables 7 and 8). Table 14 shows the detailed comparison.

The best performance was achieved by [30] 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 [27] and [28] which achieved accuracies of 0.83 and 0.81, respectively. These two works used the arXiv dataset proposed by [22] 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. Our work achieved accuracy of 0.73. Therefore, considering the abovementioned issues, this result was competitive since our model 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 our best accuracy 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 our study and that of [15] in which we have developed a predictor containing a labeling scheme of the author's intentions to predict the paper quality. The difference is that while [15] used the author's intentions in the Related Work section, which may cover both *citing* and *regular sentences*, our study 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, we 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 [15] showed the best accuracy of 0.7 in the poor-average-good task. These findings indicate that our *citation functions* labeling scheme is more effective than the intention labels proposed in [15]. Additionally, covering the author's intention in the entire section of this paper is crucial to assess the paper's quality rather than only in the Related Work section.

B. PERFORMANCE OF REGRESSION TASKS

This section presents the regression task experiment results for predicting the average review score (Table 15), the individual review score (Table 16), and the top 10 most influential features in both regression tasks (Table 17).

TABLE 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 paper 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.

TABLE 14. The accuracy-focused performance comparison between this paper and previous works.

Existing Works	Dataset	The Best Accuracy
[22]	arXiv.CS	0.79
[11]	ICLR 2017	0.78#
[25]	ICLR 2017	0.71
[24]	ICLR 2017–2018	0.71#
[26]	ICLR 2017	0.65
[27]	arXiv and ICLR	0.83
[28]	arXiv.CS.CL	0.81
[29]	arXiv and ICLR	0.78#
[30]	ICLR 2017	0.85
[23]	AI conference, Robotics	0.77#
[12]	ICLR 2017	0.88#
[33]	ICLR 2019	0.77#
[34]	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. We noted that studies by [31] and [32] used f1 and AUC, respectively. Since the best results of these works include the arXiv dataset obtained from [22], which determined the acceptance status as accepted and "probably-rejected," we prefer not to put the results in this Table. Moreover, the comparison is also inapplicable to the study by [15] because they only focused on the Related Work section, annotators rather than the editor determined the target classes.

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

TABLE 15. The best performance of average review score prediction for each regression scenario. The bold values indicate the lowest result achieved by each algorithm.

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

TABLE 16. The best performance of individual review score prediction for each regression scenario. The bold values indicate the lowest result achieved by each algorithm.

Predictors	Individual review Score	RFR		GBR		SVR		XGBR		DTR	
		n	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*	*n*	*Value*
Citing Sentence	RMSE 1	15	1.82	10	1.80	11	1.81	10	1.80	1	1.84
Predictor	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	RMSE 1	17	1.83	2	1.81	19	1.81	1	1.81	1	1.86
Predictor	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	RMSE 1	1	1.79	17	1.76	17	1.78	2	1.77	1	1.79
Predictor	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	RMSE 1	53	**1.71**	6	**1.72**	3	**1.73**	5	**1.73**	1	**1.79**
Predictor	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

TABLE 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

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.

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.

We compare the best results of regression tasks in this paper 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 [23] developed based on review comments that achieved the best RMSE and MAE of 1.28 and 1.05, respectively, which are slightly higher than our performances. However, our best performances (RMSE: 1.34, MAE: 1.07) are considered competitive since the regression method was developed based on the paper without review comments.

VI. CONCLUSION AND FUTURE WORK

This paper 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. Our prediction method 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 our hypothesis 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 our 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 our method to other domains, e.g., broader CS and medicine, among others. Second, we intend 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, we hope to be one step closer to incorporating TAPR into the entire peer-review process.

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

[1] F. Rowland, "The peer-review process," *Learned Publishing*, vol. 15, no. 4, pp. 247–258, Oct. 2002, doi: 10.1087/095315102760319206.
[2] R. Johnson, A. Watkinson, and M. Mabe. (Oct. 2018). *The STM Report—An Overview of Scientific and Scholarly Publishing*. The Hague, The Netherlands. [Online]. Available: https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf
[3] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi, "AI-assisted peer review," *Humanities Social Sci. Commun.*, vol. 8, no. 1, pp. 1–11, Dec. 2021, doi: 10.1057/s41599-020-00703-8.
[4] M. Jubb, "Peer review: The current landscape and future trends," *Learned Publishing*, vol. 29, no. 1, pp. 13–21, Jan. 2016, doi: 10.1002/leap.1008.
[5] Z. Tong, Y. Huan, S. Lei, W. Jing, and X. Daojia, "Application and classification of artificial intelligence-assisted academic peer review," *Chin. J. Sci. Tech. Periodals*, vol. 32, no. 1, pp. 65–74, 2021, doi: 10.11946/cjstp.201911220799.
[6] J. P. Tennant, "The state of the art in peer review," *FEMS Microbiol. Lett.*, vol. 365, no. 19, pp. 1–10, Oct. 2018, doi: 10.1093/femsle/fny204.
[7] S. Schroter, N. Black, S. Evans, J. Carpenter, F. Godlee, and R. Smith, "Effects of training on quality of peer review: Randomised controlled trial," *Brit. Med. J.*, vol. 328, no. 7441, pp. 673–675, Mar. 2004, doi: 10.1136/bmj.38023.700775.ae.
[8] C. A. Pierson, "Peer review and journal quality," *J. Amer. Assoc. Nurse Practitioners*, vol. 30, no. 1, pp. 1–2, Jan. 2018, doi: 10.1097/JXX.0000000000000018.
[9] D. Moher et al., "Core competencies for scientific editors of biomedical journals: Consensus statement," *BMC Med.*, vol. 15, no. 1, pp. 1–10, Sep. 2017, doi: 10.1186/s12916-017-0927-0.

[10] S. Jana, "A history and development of peer-review process," *Ann. Library Inf. Stud.*, vol. 66, no. 4, pp. 152–162, 2019.

[11] K. Wang and X. Wan, "Sentiment analysis of peer review texts for scholarly papers," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 175–184, doi: 10.1145/3209978.3210056.

[12] P. Fytas, G. Rizos, and L. Specia, "What makes a scientific paper be accepted for publication?" in *Proc. 1st Workshop Causal Inference NLP*, 2021, pp. 44–60, doi: 10.18653/v1/2021.cinlp-1.4.

[13] R. L. Kravitz, P. Franks, M. D. Feldman, M. Gerrity, C. Byrne, and W. M. Tierney, "Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care?" *PLoS ONE*, vol. 5, no. 4, pp. 2–6, 2010, doi: 10.1371/journal.pone.0010072.

[14] A. S. Raamkumar, S. Foo, and N. Pang, "Survey on inadequate and omitted citations in manuscripts: A precursory study in identification of tasks for a literature review and manuscript writing assistive system," *Inf. Res.*, vol. 21, no. 4, pp. 1–30, 2016, [Online]. Available: http://informationr.net/ir/21-4/paper733.html

[15] A. J. Casey, B. Webber, and D. Glowacka, "Can models of author intention support quality assessment of content?" in *Proc. Bibliometric-Enhanced Inf. Retr. Natural Lang. Process. Digit. Libraries (BIRNDL 2019)*, 2019, pp. 92–99. [Online]. Available: http://ceur-ws.org/Vol-2414/

[16] K. L. Lin and S. X. Sui, "Citation functions in the opening phase of research articles: A corpus-based comparative study," in *Corpus-based Approaches to Grammar, Media and Health Discourses* (The M.A.K. Halliday Library Functional Linguistics Series). Singapore: Springer, 2020, pp. 233–250, doi: 10.1007/978-981-15-4771-3_10.

[17] I. Tahamtan and L. Bornmann, "What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018," *Scientometrics*, vol. 121, no. 3, pp. 1635–1684, Dec. 2019, doi: 10.1007/s11192-019-03243-4.

[18] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles' metadata and cue-terms from content," *Scientometrics*, vol. 118, no. 1, pp. 21–43, Jan. 2019, doi: 10.1007/s11192-018-2961-x.

[19] M. Roman, A. Shahid, S. Khan, L. Yu, M. Asif, and Y. Y. Ghadi, "Investigating maps of science using contextual proximity of citations based on deep contextualized word representation," *IEEE Access*, vol. 10, pp. 31397–31419, 2022, doi: 10.1109/ACCESS.2022.3159980.

[20] S. Basuki and M. Tsuchiya, "SDCF: Semi-automatically structured dataset of citation functions," *Scientometrics*, vol. 127, no. 8, pp. 4569–4608, Aug. 2022, doi: 10.1007/s11192-022-04471-x.

[21] W. Yuan, P. Liu, and G. Neubig, "Can we automate scientific reviewing?" *J. Artif. Intell. Res.*, vol. 75, pp. 171–212, Sep. 2022, doi: 10.1613/jair.1.12862.

[22] D. Kang, W. Ammar, B. Dalvi, M. Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (PeerRead): Collection, insights and NLP applications," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1647–1661, doi: 10.18653/v1/N18-1149.

[23] A. C. Ribeiro, A. Sizo, H. L. Cardoso, and L. P. Reis, "Acceptance decision prediction in peer-review through sentiment analysis," in *Progress in Artificial Intelligence* (Lecture Notes in Artificial Intelligence), vol. 12981. Cham, Switzerland: Springer, 2021, pp. 766–777, doi: 10.1007/978-3-030-86230-5_60.

[24] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, "DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1120–1130, doi: 10.18653/v1/P19-1106.

[25] W. Jen and M. Chen, "Predicting conference paper acceptance," Stanford Univ., Stanford, CA, USA, Mach. Learn. Course Project 117, 2018. [Online]. Available: https://cs229.stanford.edu/proj2018/report/117.pdf

[26] A. Ghosh, N. Pande, R. Goel, R. Mujumdar, and S. S. Sistla. (2020). *Prediction, Conference Paper Acceptance (Acceptometer)*. Atlanta, GA, USA. [Online]. Available: https://rohangoel.com/Acceptometer/;

[27] M. Skorikov and S. Momen, "Machine learning approach to predicting the acceptance of academic papers," in *Proc. IEEE Int. Conf. Ind., Artif. Intell., Commun. Technol. (IAICT)*, Jul. 2020, pp. 113–117, doi: 10.1109/IAICT50021.2020.9172011.

[28] G. M. de Buy Wenniger, T. van Dongen, E. Aedmaa, H. T. Kruitbosch, E. A. Valentijn, and L. Schomaker, "Structure-tags improve text classification for scholarly document quality prediction," in *Proc. 1st Workshop Scholarly Document Process.*, 2020, pp. 158–167, doi: 10.18653/v1/2020.sdp-1.18.

[29] A. Ciloglu and M. Merdan, "Big peer review challenge," Seminar Inf. Syst. (WS19/20), Humboldt-Universität, Berlin, Germany, 2020. Accessed: Dec. 4, 2022. [Online]. Available: https://humboldt-wi.github.io/blog/research/information_systems_1920/group11_peer_reviews/

[30] D. J. Joshi, A. Kulkarni, R. Pande, I. Kulkarni, S. Patil, and N. Saini, "Conference paper acceptance prediction: Using machine learning," in *Machine Learning and Information Processing*. Singapore: Springer, 2021, pp. 143–152, doi: 10.1007/978-981-33-4859-2_14.

[31] P. Vincent-Lamarre and V. Larivière, "Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome," *Quant. Sci. Stud.*, vol. 2, no. 2, pp. 662–677, Jul. 2021, doi: 10.1162/qss_a_00125.

[32] P. Bao, W. Hong, and X. Li, "Predicting paper acceptance via interpretable decision sets," in *Proc. Companion Web Conf.*, Apr. 2021, pp. 461–467, doi: 10.1145/3442442.3451370.

[33] P. K. Bharti, S. Ranjan, T. Ghosal, M. Agrawal, and A. Ekbal, "PEERAssist?: Leveraging on paper-review interactions to predict peer review decisions," in *Proc. Int. Conf. Asian Digit. Libraries*, vol. 1, pp. 421–435, Jan. 2021, doi: 10.1007/978-3-030-91669-5.

[34] T. Pradhan, C. Bhatia, P. Kumar, and S. Pal, "A deep neural architecture based meta-review generation and final decision prediction of a scholarly article," *Neurocomputing*, vol. 428, pp. 218–238, Mar. 2021, doi: 10.1016/j.neucom.2020.11.004.

[35] L. Stappen, G. Rizos, M. Hasan, T. Hain, and B. W. Schuller, "Uncertainty-aware machine support for paper reviewing on the Interspeech 2019 submission corpus," in *Proc. Interspeech*, Oct. 2020, pp. 1808–1812, doi: 10.21437/Interspeech.2020-2862.

[36] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, and N. F. Rajani, "ReviewRobot: Explainable paper review generation based on knowledge synthesis," in *Proc. 13th Int. Conf. Natural Lang. Gener.*, 2020, pp. 384–397. [Online]. Available: https://aclanthology.org/2020.inlg-1.44

[37] J. Li, A. Sato, K. Shimura, and F. Fukumoto, "Multi-task peer-review score prediction," in *Proc. 1st Workshop Scholarly Document Process.*, 2020, pp. 121–126, doi: 10.18653/v1/2020.sdp-1.14.

[38] M. Färber, A. Thiemann, and A. Jatowt, "A high-quality gold standard for citation-based tasks," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 1885–1889. [Online]. Available: https://www.aclweb.org/anthology/L18-1296

[39] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3615–3620, doi: 10.18653/v1/D19-1371.

SETIO BASUKI received the bachelor's degree from STT Telkom (Telkom University), in 2007, and the master's degree from the Institute Teknologi Bandung (ITB), Indonesia, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Toyohashi University of Technology (TUT), Japan. Since 2009, he has been with the Informatics Study Program, Universitas Muhammadiyah Malang, Indonesia, as a Faculty Member. His research interests include machine learning and natural language processing.

MASATOSHI TSUCHIYA received the B.E., M.E., and Dr. degrees in informatics from Kyoto University, Kyoto, Japan, in 1998, 2000, and 2007, respectively. In 2004, he joined the Computer Center, Toyohashi University of Technology, Toyohashi, Japan, as an Assistant Professor. His section was re-constructed to the Information and Media Center, in 2005. Since 2014, he has been an Associate Professor with the Toyohashi University of Technology. His major research interest includes natural language processing.