Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 9, No. 4, November, Pp. 409-420

Predicting the sentiment of review aspects in the peer review text using machine learning

Setio Basuki*[1], Zamah Sari[1], Masatoshi Tsuchiya[2], Rizky Indrabayu[1]
Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia[1]
Computer Science and Engineering, Toyohashi University of Technology, Japan[2]

Article Info
Keywords:
Citation Functions, Paper Quality, Peer Review, Review Aspects, Sentiment Analysis

Article history:
Received: June 08, 2024
Accepted: August 01, 2024
Published: November 30, 2024

Cite:
S. Basuki, Z. Sari, M. . Tsuchiya, and R. Indrabayu, "Predicting the Sentiment of Review Aspects in the Peer Review Text using Machine Learning", *KINETIK*, vol. 9, no. 4, Nov. 2024.
https://doi.org/10.22219/kinetik.v9i4.2042

*Corresponding author.
Setio Basuki
E-mail address:
setio_basuki@umm.ac.id

Abstract
This paper develops a Machine Learning (ML) model to classify the sentiment of review aspects in the peer review text. Reviewers use the review aspect as paper quality indicators such as motivation, originality, clarity, soundness, substance, replicability, meaningful comparison, and summary during the review process. The proposed model addresses the critique of the existing peer review process, including a high volume of submitted papers, limited reviewers, and reviewer bias. This paper uses citation functions, representing the author's motivation to cite previous research, as the main predictor. Specifically, the predictor comprises citing sentence features representing the scheme of citation functions, regular sentence features representing the scheme of citation functions for non-citation sentences, and reference-based representing the source of citation. This paper utilizes the paper dataset from the International Conference on Learning Representations (ICLR) 2017-2020, which includes sentiment values (positive or negative) for all review aspects. Our experiment on combining XGBoost, oversampling, and hyper-parameter optimization revealed that not all review aspects can be effectively estimated by the ML model. The highest results were achieved when predicting Replicability sentiment with 97.74% accuracy. It also demonstrated accuracies of 94.03% for Motivation and 93.93% for Meaningful Comparison. However, the model exhibited lower effectiveness on Originality and Substance (85.21% and 79.94%) and performed less effectively on Clarity and Soundness with accuracies of 61.22% and 61.11%, respectively. The combination predictor was the best for the 5 review aspects, while the other 2 aspects were effectively estimated by regular sentence and reference-based predictors.

1. Introduction

Assessing scientific paper quality through peer review has become a widespread standard in the academic communities for journal publishing, conference submission, and grant assessment [1]. Finishing the peer review process is a time- and energy-consuming task, starting from receiving manuscripts to reaching a final decision. Peer review poses a challenge during the process due to the exponential increase in manuscript submissions. According to the STM report in 2018 [2], over 3 million research articles are published annually (there are 33,100 journals in English and 9,400 journals in languages other than English). Based on another report, the annual time spent for reviewing articles that were previously rejected amounted to 15 million hours [3]. Additionally, EasyChair, a conference management platform, has handled more than 100,000 conferences since 2002. Moreover, the uneven geographic distribution of reviewers brings another burden on the overall peer review system [4].

Another challenge in the existing peer review process is the bias caused by various factors including reviewers' experience, emotions, and academic background [5]. The next challenges in this process, as highlighted by [6], include a lack of proper guidance on how to manage peer review [7], an unequal relation between journal quality and peer review [8], and the absence of competency standards for editors [9]. Furthermore, Jana [10] adds several issues, including the expensive and slow publication time of the review process, and unethical comments from, reviewers who are not willing to finish all stages of the review process. Hence, this scenario offers a challenge for the development of TAPR to handle the peer review workload.

The research on Technology-assisted Peer Review (TAPR) has obtained a lot of focus as an innovative way to reduce the review workload, particularly in three main aspects: predicting the quality of scientific articles, final editor decision, and review scores. TAPR has been designed for several purposes, from predicting three review decisions—accepted, borderline, or rejected [11], [12]—to predicting two types of results—accepted or rejected [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. However, studies regarding this topic indicated two major drawbacks including inconsistency between the review results and the final decisions, as reported by [26], and the use of the review

results as the basis for quality prediction. This approach is considered as impractical and unfair when the primary goal of TAPR is to solve the review workload.

Predicting the sentiment of review aspects is crucial, as it acts as a measure of the quality of the paper. The review aspects are the paper quality indicators used for evaluating the submitted paper during peer review including motivation, originality, clarity, soundness, substance, replicability, meaningful comparison, and summary (see Table 1). Several researchers have undertaken the task of estimating the sentiment in peer review texts. The works started from extracting the peer review text and its corresponding sentiment [27], extracting the fine-grained aspects and its scores [28], [29], predicting the score of the review aspects [15], and predicting the paper acceptance based on the review aspects [30], [31]. The majority of these works aim to predict the sentiment (positive/negative) of peer review texts that represent the reviewers' sentiments on all review aspects. The use of the review text as the basis of prediction is considered impractical to predict the paper quality since the TAPR aims to reduce the workload by directly extracting prediction features from the paper.

This paper develops a machine learning model to predict the sentiment of review aspects with the paper's content as the basis of prediction. The model is created using *citation functions* that have been developed in our previous research [32] that can be divided into more specific features, i.e., *citing sentence* features representing the scheme of *citation functions*, *regular sentence* features representing the scheme of citation functions for non-citation sentence, and *reference-based* representing the source of citation. Several reasons for using the citation function for paper quality assessment are identifying the positioning of the proposed research in the broad literature [33], understanding the comprehensive view of certain research topics [34], indicating the novelty of the proposed research [35], and estimating the research quality [12]. Moreover, the *citation functions*-based predictor demonstrated its effectiveness in predicting the final result (accepted/rejected), paper quality (good/poor), and review scores as shown in our work [36]. Lastly, there were no studies that used the citation function to predict the sentiment of the review aspect.

Here, the prediction model is developed using eXtreme Gradient Boosting (XGBoost) and is accompanied by feature selection, data balancing methods, and hyperparameter optimization. The dataset used in this paper was obtained from the International Conference on Learning Representations (ICLR) 2017-2020 which provides the dataset and its corresponding review results. This paper provided several contributions as follows:

- This paper introduces three predictors, namely the *citing sentence* feature, *regular sentence* feature, and *reference-based* feature, inspired by *citation functions*. These predictors aim to predict the sentiment of review aspects within the review text.
- The highest performance of 97.74% to estimate the sentiment on the *Replicability* aspect was achieved through a combination of XGBoost, oversampling, and hyper-parameter optimization.
- Analyzing the most important features indicates the superiority of the *reference-based* predictor, which had 89 occurrences, while the *citing sentence* and *regular sentence* predictors exhibited comparable results with 65 and 83 occurrences, respectively.
- In summary, the *citation functions* representing the motivation of authors to cite previous work are effective in predicting the sentiment of the review aspects.

2. Sentiment Analysis Methods

This section shows how the proposed prediction system for classifying the sentiment of review aspects is developed. The proposed system is developed using citation functions representing the reason why authors of research papers cite previous works. Here, this paper divides the whole system into several stages. In the first stage, we focus on obtaining the paper dataset which is accompanied by the peer review results containing the review text, review score, and sentiment of review aspects. Following this, the second stage focuses on extracting the citation functions from papers and the aspect sentiment from corresponding peer review results. In the third stage, this paper develops machine learning models based on citation functions to predict the sentiment of the review aspects. Finally, the fourth stage evaluates the performance of the prediction model that has been developed in the previous stage. Furthermore, the final stage presents the analysis of what features are significant for the prediction, and the relationship between the sentiment and the final decision (accepted or rejected), paper quality (good or poor), and reviewer score.

There are several important technical terms used in this paper, namely citing paper which means that the paper is citing other/previous papers, cited paper is the paper that is cited by the citing paper, citing sentence is the sentence containing citation mark, and regular sentence is the sentence without citation mark. The illustration of these technical terms is shown in Figure 1, and the system architecture for developing the prediction system is depicted in Figure 2.

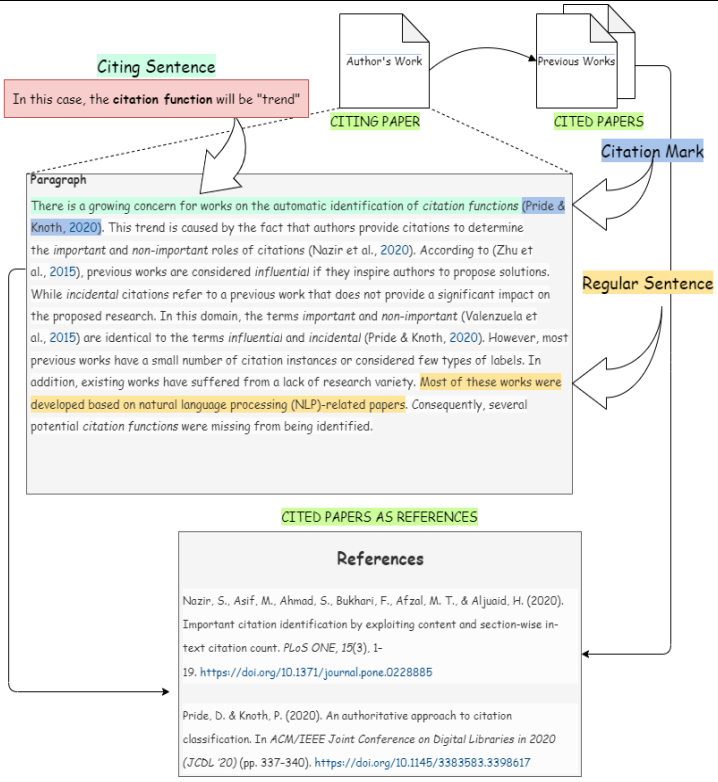

Figure 1. Several Technical Terms Used in this Paper: Citation Mark, Citing Paper, Cited Paper, Citing Sentence, and Regular Sentence

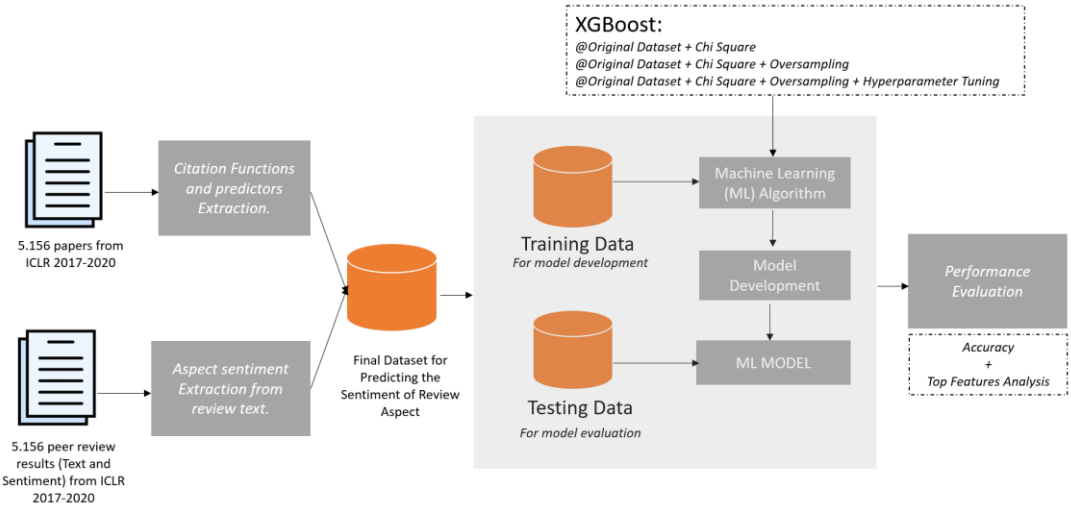

Figure 2. The System Architecture for Predicting the Sentiment of Review Aspects

2.1 Review Aspects during Peer Review Process

Generally, the peer review process involves an editor and two or more reviewers. The final decision of paper acceptance will be determined by the editor by considering the comments made by reviewers. In the International Conference on Learning Representations (ICLR), the reviewers are supported with review guidance to verify and assess eight aspects representing the quality of the papers. The reviewers are encouraged to give their comments in three forms: (a) review text, (b) review score, and (c) review aspect sentiment. This paper focuses on the aspect sentiment when reviewers judge the eight review aspects with positive or negative. The detail of the eight review aspects used in ICLR is shown in Table 1.

Table 1. Eight Review Aspect with Explanations

Review Aspects	Explanations
Motivation/Impact	Evaluating the motivation, idea, and potential impact/significance of the paper.
Originality/Novelty	Evaluating the novelty and originality of the paper.
Clarity	Evaluating the writing accuracy and clarity of the paper.
Soundness	Evaluating the quality of the analysis, design, and experiment results of the paper.
Substance	Evaluating the complexity, depth, and contribution of the paper.
Replicability	Evaluating the experiment replicability of the paper.
Meaningful Comparison	Evaluating the quality of comparing and contrasting aspects in the paper.
Summary	Evaluating the quality of the paper summary.

2.2 Paper and Peer Review Dataset

This paper uses a dataset of papers and their review results from ICLR 2017-2020, containing 5,156 papers [37], as shown in Table 2. While the final decision of paper acceptance is already available in the dataset and is determined by the ICLR's editor, the paper quality (good/poor) is adapted from the research conducted in our previous research [36].

The sentiment of review aspects is available as part of the review results. Here, we parse the values and arrange them as the target prediction. However, not all papers in our dataset are accompanied by the aspect's values. For example, the values are not available for one or more aspects, but there are cases in which the vales are unavailable for all aspects. As a result, the dataset used for the prediction is smaller than the total number of papers, as depicted in Table 2.

Table 2. Distribution of ICLR Papers Used in the Prediction System

Year-	Paper Status				
	Accept	Reject	Good	Poor	Total
ICLR-2017	198	289	416	71	487
ICLR-2018	336	571	769	138	907
ICLR-2019	502	1048	1275	275	1550
ICLR-2020	686	1526	1115	1097	2212
ALL	1722	3434	3575	1581	5156

Since this paper focuses on predicting the sentiment of review aspects of the accepted papers, the quantity of papers used in our experiments is fewer than those described in Table 2. The challenge in preparing the final dataset is that each paper has one to three sentiments for each aspect. This is because each paper was being reviewed by more than one reviewer. To handle this, we only employ papers that have consistent sentiment. For example, papers that obtained two similar sentiments (both positive or both negative) for the aspect of Clarity will be used, but papers that received different sentiments will be eliminated. Here, the sentiment values of the aspect summary are not reported, since it is unavailable on the dataset. The final dataset distribution is shown in Table 3.

Table 3. The Positive-Negative Sentiment Distribution of Each Review Aspect

Review Aspects	Positive	Negative	Total
Motivation/Impact	734	150	884
Originality	571	291	862
Clarity	425	447	872
Soundness	447	446	893
Substance	286	560	846
Replicability	34	471	505
Meaningful Comparison	118	595	713
Summary	-	-	-

2.3 Citing Sentence Features

The main classification feature used in this paper is *citing sentences* representing sentences that contain citation marks. This feature is created by extracting and categorizing *citing sentences* from all papers in the dataset into 18 labels of *citation functions*. The categorization process is performed using the Bidirectional Encoder Representations from Transformers (BERT)-based model developed by [32]. The final features are generated by calculating the presence of each feature's label in each paper.

We denote this feature from (c-0) to (c-18), and add an additional feature (c-19) representing the quantity of *citing sentence* found in the *citing paper*, as shown in Table 4.

Table 4. The 20 Labeling Scheme of Citation Functions

[Generic/Coarse Label]: Background
Explaining to the theories, principles, concepts, topics, problems, etc. stated on the cited papers.
Fine-grained Label:
• **(c-0) definition,** definition of general concept, principle, topic, or problem. *example:* Neural Machine Translation (NMT) is novel framework for text translation between languages <citation>. • **(c-1) suggest,** suggestion to check more detail, refer, or explore other papers. *Example*: The interested reader can refer to <citation> for further information. • **(c-2) judgment,** showing the positive or negative, useful or not-useful of certain topics or problems. *example*: The n-coalescent has interesting statistical properties <citation>. • **(c-3) technical,** explaining how a principle is applied. *Example*: The inference stage is implemented using Gibbs sampling technique <citation>. • **(c-4) trend,** expressing the significance of the theory, principle, concept, topic, or problem. *example*: A recent trend showing by <citation> demonstrates that deeper CNNs reach better results.
[Generic/Coarse Label]: Citing Paper Work
The author's work.
Fine-grained Label:
• **(c-5) corroboration,** the citing paper cites cited paper while proposing a research topic. *example*: We propose a Minimum Message Length technique of causal discovery <citation> in Section 4. • **(c-6) based on,** the citing paper follow or inspired by the cited paper. *Example*: we focus on the parallelism mechanism of the decoder and the consumed energy, as inspired by <citation>. • **(c-7) use,** the citing paper implement the technique, dataset, or technique from the cited paper. *example*: We use the unsupervised dependency parser (UDP) implemented by <citation>. • **(c-8) extend,** the citing paper adapt, improves, or modifies the work of cited paper. *example*: in order to make it applicable, we modify the microscopic search rules proposed by <citation>. • **(c-9) dominant,** the citing paper' performance exceeds the cited paper' performance. *Example*: our proposed method outperforms current state of the art (SoTA) on both languages <citation>. • **(c-10) future,** the citing paper' the future work. *Example*: in the future, we plan to explore the distributed variants of S3GD like <citation>.
[Generic/Coarse Label]: Cited Paper Work
The cited papers' work.
Fine-grained Label:
• **(c-11) propose,** explaining the proposed research of the cited paper. *Example*: the work by <citation> proposes a model for storing and operating on infra-red images. • **(c-12) success,** stating the accomplishment of the cited paper. *Example*: successful extraction has been done by <citation> for body appearance and topology from real and synthetic data. • **(c-13) weakness,** stating the limitation of the cited paper. *Example*: focusing only on two-user communication system is the limitation of <citation>. • **(c-14) result,** stating the experiment result of the cited paper in a neutral way. *Example*: a precision of 0.97 and a recall of 0.83 were achieved by the JavaBaker <citation>. • **(c-15) dominant,** the superiority of the cited paper' performance over the citing paper. *Example*: only the work by <citation> proposing deeper ResNet outperformed our method.
[Generic/Coarse Label]: Compare and Contrast
Comparing and contrasting between the citing paper and cited paper.
Fine-grained Label:
• **(c-16) compare,** similarity between the citing paper and cited paper. *Example*: the BLHT technique <citation> is similar to our work.

• **(c-17) contrast,** how the citing paper differ from the cited paper. *Example*: different with <citation>, our proposed model does not have a partially nested information.
[Generic/Coarse Label]: Other
For stating the citing sentences that do not meet the previous criteria
Fine-grained Label:
• **(c-18) comparison,** comparison between cited papers (can be similarities/differences). *Example*: the computational complexity comparison of the proposed method with AOG <citation> and nCTE <citation>. • **(c-18) multiple_intent,** single citing sentence consisting more than one citation marks for different intents. *Example*: the work by Blum, Luby and Rubinfeld <citation> is one of the early researches which dealt with linearity testing (refer to <citation> for testing in low degree scenario). • **(c-18) other,** citing sentences that do not meet all of the label definition. *Example*: The first paper is by Sethuraman and Sab'an <citation>.
[Generic/Coarse Label]: Additional
• **(c-19) num_citing_sent.,** storing the number of citing sentences.

2.4 Regular Sentence Features

Adopting the concept of *citing sentence* features, the *regular sentence* features are generated by classifying the *regular sentence* using the BERT model into 18 labels of *citation functions* as shown in Table 4. The final features are formed by computing the presence of each label in each paper in the dataset. The reason we propose this feature is that an author's intention to cite previous works cannot always be captured using only citing sentences. We denote this feature from (r-20) to (r-38) and add an additional feature (r-39) representing the number of regular sentences.

2.5 Reference-based Features

This paper proposes the *reference-based* features as an additional classification predictor. The aim of this feature is to clarify the impact of the references (source of citation) in the sentiment classification. This feature contains of 24 labels which can be divided into several groups, i.e., generic, preprint, conference, and journal, as shown in Table 5. To generate this feature, we extract all reference sections of each paper and calculate the presence of the label using a rule-based approach. The features are denoted with (ref-0) to (ref-23).

Table 5. The Reference-based Features Consisting of 24 Labels

Common Labels
• **(ref-0)** Total references in the paper **(NUM_REF)** • **(ref-1)** Total references published within 3 years (**NUM_REF_3YEARS)**
Preprint Labels
• **(ref-2)** Open Access Repository (preprint) (**arXiv**)
Top Conference Labels
• **(ref-3)** Conference_on_Neural_Information_Processing_Systems **(NeurIPS)** • **(ref-4)** International_Conference_on_Learning_Representations **(ICLR**) • **(ref-5)** International_Conference_on_Machine_Learning **(ICML)** • **(ref-6)** Association_for_the_Advancement_of_Artificial_Intelligence **(AAAI**) • **(ref-7)** International_Conference_on_Computer_Vision **(ICCV)** • **(ref-8)** Conference_on_Computer_Vision_and_Pattern_Recognition **(CVPR)** • **(ref-9)** Empirical_Methods_in_Natural_Language_Processing **(EMNLP**) • **(ref-10)** Association_for_Computational_Linguistics (**ACL)** • **(ref-11)** North_American_Chapter_of_the_Association_for_Computational_Linguistics (**NAACL**) • **(ref-12)** European_Conference_on_Computer_Vision (**ECCV**) • **(ref-13)** The International Conference on Robotics and Automation **(ICRA**) • **(ref-14)** the_International_Conference_on_Acoustics_Speech_and_Signal Processing **(ICASSP**) • **(ref-15)** The_International_Joint_Conference_on_Artificial_Intelligence **(IJCAI**) • **(ref-16)** The_International_Conference_on_Artificial_Intelligence_and_Statistics **(AISTATS**) • **(ref-17)** Special_Interest_Group_on_Knowledge_Discovery_and_Data_Mining **(SIGKDD**)
Popular Journal Labels
• **(ref-18)** Neural Computation **(Neuralcom**) • **(ref-19) IEEE Transaction** • **(ref-20) ACM Transaction**

- **(ref-21) MIT Press**
- **(ref-22) Nature**
- **(ref-23)** The Journal of Machine Learning Research **(JMLR**)

2.6 The Experiment Scenario of Sentiment Analysis

The prediction system is seen as a classification problem with two target sentiment classes: positive and negative. The classification experiments are performed based on four scenarios using: (i) citing sentence predictor, (ii) regular sentence predictor, (iii) reference-based predictor, and (iv) combination predictor. Each scenario involves three prediction settings: original dataset with feature selection, balanced dataset with feature selection, and balanced dataset with feature selection and hyper-parameter optimization. This paper uses Chi-Square for feature selection and random oversampling for data balancing. Since the format of the generated features is tabular, XGBoost is the most appropriate classification algorithm due to its superior performance in many experiments [38]. We look for the appropriate values for three types of hyperparameters, including learning rate, number of estimators, and maximum depth. Notably, all scenarios will be applied to all review aspects independently, and the classification performance of each paper is measured using accuracy and the list of features for obtaining the best outcomes.

3. Results and Discussion

This section shows the results of our experiments on the prediction of sentiment of review aspects. There are three types of results that will be reported: (1) the comparison of classification performance, (2) an analysis of the most significant features for achieving the best performance, and (3) a discussion of the relationship between the review sentiment, the acceptance decision, paper quality, and review score.

3.1 The Classification Performances

Table 6 displays the best outcomes for each classification scenario. For all review aspects, the lowest performances were obtained using the original dataset even though it has been combined with the chi-square. There is a significant increase in accuracy when trying to make the original dataset more balanced through oversampling, and the best performances were reached by applying the hyper-parameter tuning on XGBoost. The ML model demonstrates its effectiveness in predicting the reviewer's sentiment on Replicability, achieving an accuracy of 97.74% using the regular sentence predictor with 17 attributes. The model performs well to predict Meaningful Comparison and Motivation, with accuracies of 93.93% and 94.03%, respectively. Subsequently, it relatively predicts the sentiment on Originality and Substance with accuracies of 85.21% and 79.94%. However, the model is less effective in predicting the reviewer's sentiment on Clarity (61.22%) and Soundness (61.11%).

The high accuracies for several review aspects indicate that the characteristics of these aspects can be effectively represented by the prediction features. For example, the meaningful comparison can be easily recognized from sentences in the paper corresponding to features that compare citing and cited papers, including "compare", "contrast", "based on", "use", "extend", and "dominant" found in both *citing sentences* and *regular sentences*.

The low prediction accuracy values for aspects such as clarity and soundness indicate that the citation function-based feature is ineffective in predicting the aspects' sentiment. For example, in the clarity aspect, the prediction feature fails to represent the quality of the writing in the paper—whether the paper is easy to follow, well-organized, clear, and concise. To assess this, the reviewer needs to evaluate the entire text and cannot rely solely on sentence fragments from citing or regular sentences.

Table 6. The Best Accuracy for Each Prediction Scenario

Aspect Review	Predictors	Original Data + Chi Square		Oversampling + Chi Square		Oversampling + Chi Square + Hyperparameter Tuning	
		N	Acc. (%)	N	Acc. (%)	N	Acc. (%)
Clarity	citing sentence	12	57.25	7	55.51	18	57.79
	regular sentence	1	55.29	20	57.03	19	54.37
	reference-based	9	56.08	6	58.94	11	59.70
	combination	63	60.78	41	**61.22**	46	60.46
Meaningful Comparison	citing sentence	2	87.98	15	92.01	20	92.97
	regular sentence	12	86.34	10	92.97	10	93.61
	reference-based	7	85.25	15	92.01	15	92.01
	combination	62	86.34	58	**93.93**	62	93.61

Motivation	citing sentence	1	85.31	17	92.36	20	92.84
	regular sentence	1	85.31	19	93.08	13	**94.03**
	reference-based	1	84.90	18	91.65	14	92.60
	combination	24	85.31	44	93.08	41	93.79
Originality	citing sentence	1	67.08	18	76.92	15	78.11
	regular sentence	19	65.02	13	79.29	13	80.77
	reference-based	6	69.55	15	76.92	15	77.81
	combination	23	71.19	43	83.73	22	**85.21**
Replicability	citing sentence	6	94.07	20	96.38	14	97.29
	regular sentence	20	94.07	19	96.38	18	97.29
	reference-based	18	94.07	21	96.38	20	96.38
	combination	64	94.07	49	97.29	26	**97.74**
Soundness	citing sentence	2	57.83	2	54.76	20	54.76
	regular sentence	5	52.21	15	49.21	20	51.59
	reference-based	16	59.44	17	58.33	16	**61.11**
	combination	45	57.43	56	55.56	38	57.54
Substance	citing sentence	1	67.10	15	77.12	15	79.62
	regular sentence	1	72.73	19	77.43	14	79.31
	reference-based	1	70.56	19	76.18	20	79.62
	combination	1	70.56	24	78.06	61	**79.94**

Table 7 shows the hyper-parameters used by XGBoost to reach the best results on each review aspect. The parameters experimented in this research were learning rate, number of estimators, and max depth.

Table 7. The Hyper-parameter Setting for Best Scenario of Each Review Aspect

Review Aspect	Hyperparameter
Clarity	*learning_rate*=0.25, *n_estimators*=90, max_depth=10
Meaningful Comparison	*learning_rate*=0.3, *n_estimators*=50, max_depth=8
Motivation	*learning_rate*=0.30, *n_estimators*=80, max_depth=10
Originality	*learning_rate*=0.3, *n_estimators*=50, max_depth=6
Replicability	*learning_rate*=0.3, *n_estimators*=50, *max_depth*=9
Soundness	*learning_rate*=0.3, *n_estimators*=30, *max_depth*=7
Substance	*learning_rate*=0.2, *n_estimators*=60, *max_depth*=8

3.2 Most Influential Classification Features

Among all predictors, the *combination* is the most effective method to achieve the highest accuracy in most of the review aspects, except in *motivation* and *soundness* where the best results were achieved using *regular sentence* and *reference-based* predictors. Looking closely at the features, we show their distribution in Table 8. For the convenient presentation, we denote each predictor as follows: (#1) for *citing sentence*, (#2) for *regular sentence*, and (#3) for *reference-based*. This mark helps to distinguish between the *citing sentence* and the *regular sentence* predictor because they share the same list of features. Moreover, an additional marker, (citing) or (cited), was used to distinguish whether the *citing paper* outperformed the *cited paper* or vice versa.

Table 8. The Distribution of Predictor to Reach the Best performance

Predictor	Distribution
citing sentence	0
regular sentence	1
reference-based	1
combination	5

Table 9. The Feature Distribution to Reach the Highest Prediction Results

Predictor	Distribution
citing sentence	65
regular sentence	83
reference-based	89

The experiments have demonstrated other interesting results as shown in Table 8. This Table presents the feature distribution belonging to each predictor summarized from Table 9. The *reference-based* predictor was used 89 times, showing its dominance. The *regular sentence* and *citing sentence* predictors show a comparable presence of 83 and 65 respectively. Note that, the *combination* predictor is not discussed in Table 10 since it combines three other predictors.

Table 10. The Feature List to Obtain the Best Performance

Clarity	Meaningful Comparison	Motivation	Originality	Replicability	Soundness	Substance
combination	combination	regular sentence	combination	combination	reference-based	combination
#2-other	#2-corroboration	#2-judgment	#2-num_citing_sent	#2-num_citing_sent	#3-eccv	#3-num_ref_3years
#2-num_citing_sent,	#2-num_citing_sent	#2-corroboration	#2-other	#2-weakness	#3-icassp	#1-num_citing_sent
#3-neurips	#3-num_ref	#2-success	#2-corroboration	#2-judgment	#3-emnlp	#3-num_ref
#2-technical	#3-num_ref_3years	#2-result	#1-num_citing_sent	#2-success	#3-acl	#3-neurips
#3-num_ref_3years	#3-acm_tran	#2-num_citing_sent	#1-success	#1-weakness	#3-nature	#3-cvpr
#3-ijca	#3-cvpr	#2-contrast	#3-acl	#3-num_ref_3years	#3-naacl	#2-dominant (citing)
#1-suggest	#1-result	#2-definition	#2-definition	#3-aaai	#3-iccv	#1-based on
#2-based on,	#3-icml	#2-propose	#2-technical	#1-technical	#3-cvpr	#2-success
#3-icra	#3-arxiv	#2-dominant (cited)	#1-based on	#1-extend	#3-icra	#1-use
#3-emnlp	#1-use	#2-compare	#3-naacl	#2-technical	#3-icml	#2-judgment
#3-iclr	#2-definition	#2-extend	#3-ijcai	#2-corroboration	#3-iclr	#3-arxiv
#1-use	#3-sigkdd	#2-based on	#2-use	#3-cvpr	#3-ijcai	#2-use
#1-num_citing_sent	#3-neurips	#2-future	#1-dominant (citing)	#2-other	#3-aaai	#2-num_citing_sent
#2-dominant (citing)	#2-based on		#1-use	#3-ijcai	#3-mit_press	#2-dominant (cited)
#3-acl	#3-icassp		#2-success	#3-aistats	#3-num_ref	#2-based on
#1-success	#2-weakness		#2-compare	#3-eccv	#3-sigkdd	#3-eccv
#1-dominant (citing)	#1-dominant (citing)		#3-'jmlr'	#3-emnlp		#3-emnlp
#3-icassp	#3-eccv		#1-propose	#3-icra		#1-other
#2-result	#3-icra		#1-dominant (cited)	#2-contrast		#1-judgment
#1-trend	#3-iccv		#3-eccv	#1-result		#3-acl
#3-arxiv	#3-jmlr		#2-contrast	#3-num_ref		#3-icassp
#3-ieee_tran	#2-propose		#3-num_ref_3years	#1-definition		#2-other
#2-definition	#2-other			#1-based on		#3-acm_tran
#1-other	#1-compare			#3-iccv		#1-dominant (citing)
#3-acm_tran	#1-based on			#2-based on		#3-iclr
#1-weakness	#2-future			#2-future		#2-propose
#2-corroboration	#2-dominant (cited)					#1-corroboration
#1-result	#3-ijcai					#1-propose
#2-extend	#2-compare					#1-success
#2-suggest	#3-iclr					#2-result
#2-compare	#3-acl					#3-aaai
#1-extend	#1-weakness					#2-corroboration
#1-technical	#1-contrast					#3-nature
#3-mit_press	#2-result					#2-compare
#1-propose	#1-other					#2-weakness
#1-judgment	#3-emnlp					#3-naacl'
#2-dominant (cited)	#2-trend					#1-future
#3-neuralcom	#2-suggest					#1-compare
#2-future	#1-corroboration					#1-suggest
#1-future	#2-use					#3-iccv
#3-aistats	#1-technical					#2-technical
	#3-nature					#3-icml
	#2-success					#1-contrast
	#3-aistats					#1-dominant (cited)
	#1-judgment					#1-extend
	#1-extend					#1-result
	#2-judgment					#2-future
	#3-ieee_tran					#3-icra
	#1-num_citing_sent					#3-neuralcom
	#1-success					#2-extend
	#1-dominant (cited)					
	#1-future					

#3-aaai	#2-trend
#2-dominant (citing)	#1-weakness
#3-naacl	#1-definition
#1-definition	#1-trend
#2-technical	#3-ijcai
#1-suggest	#1-technical
	#2-contrast
	#3-mit_press
	#3-sigkdd
	#2-suggest
	#2-definition

The analysis of the most important feature as shown in Table 10 strengthened our hypothesis that the predictor developed based on our previous study [36] shows its effectiveness in predicting the research paper quality. In our previous work, the predictors demonstrated competitiveness in estimating the final editorial decision (acceptance or rejection) and the quality of the paper (good or poor). In this paper, we report their competitiveness in predicting the sentiment of review aspects. Since the *citation functions*-based features are extracted from the paper, it is more practical to implement the TAPR compared to several previous works mentioned in the Introduction section. Another benefit of utilizing this feature is its interpretability, which makes the proposed prediction system suitable for assisting in the evaluation of paper quality.

4. Conclusion

Predicting the sentiment of review aspects is crucial, as it serves as an early indicator to assess the paper quality. This paper has developed a machine learning model for predicting the sentiment of review aspects of the review text. The prediction model was developed using citation functions-based classification features i.e., citing sentences, regular sentences, and an additional feature called reference-based feature. Our experiments reveal that predicting the aspect sentiment of peer review text can be estimated using citation functions-based predictor. The combination of data balancing through oversampling, feature selection using chi-square, and hyper-parameter optimization delivers the best performance to predict the sentiment.

In this research, we demonstrate that the citation functions as the authors' motivation to cite previous work is useful for estimating the review sentiment. Thus, the researcher and reviewer need to consider the appropriate citations when working on a research paper. In the future, we plan to expand the scope of the paper quality into citation count prediction by using the same labeling scheme of citation functions used in this paper.

References

[1] F. Rowland, "The peer-review," *Learn. Publ.*, vol. 15, no. 4, pp. 247–258, 2002. 10.1087/095315102760319206
[2] R. Johnson, A. Watkinson, and M. Mabe, "The STM Report - An overview of scientific and scholarly publishing." Oct. 2018.
[3] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi, "AI-assisted peer review," *Humanit Soc Sci Commun*, vol. 8, no. 1, Dec. 2021. https://doi.org/10.1057/s41599-020-00703-8
[4] M. Jubb, "Peer review: The current landscape and future trends," *Learn. Publ.*, vol. 29, no. 1, pp. 13–21, 2016. https://doi.org/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," *Chinese J. Sci. Tech. Periodals*, vol. 32, no. 1, pp. 65–74, 2021. https://doi.org/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, 2018. https://doi.org/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," *Br. Med. J.*, vol. 328, no. 7441, pp. 673–675, Mar. 2004. https://doi.org/10.1136/bmj.38023.700775.AE
[8] C. A. Pierson, "Peer review and journal quality," *J. Am. Assoc. Nurse Pract.*, vol. 30, no. 1, pp. 1–2, Jan. 2018. https://doi.org/10.1097/jxx.0000000000000018
[9] D. Moher and others, "Core competencies for scientific editors of biomedical journals: Consensus statement," *BMC Med*, vol. 15, no. 1, Sep. 2017. https://doi.org/10.1186/s12916-017-0927-0
[10] S. Jana, "A history and development of peer-review process," *Ann. Libr. 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 *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, Ann Arbor Michigan, U. S. A.: Association for Computing Machinery*, 2018, pp. 175–184. https://doi.org/10.1145/3209978.3210056
[12] A. J. Casey, B. Webber, and D. Glowacka, "Can models of author intention support quality assessment of content?," in *Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019), Paris, France: CEUR Workshop Proceedings (CEUR-WS.org)*, 2019, pp. 92–99.
[13] P. Fytas, G. Rizos, and L. Specia, "What Makes a Scientific Paper be Accepted for Publication?," in *Proceedings of the First Workshop on Causal Inference and NLP, Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021*, 2021, pp. 44–60. https://doi.org/10.18653/v1/2021.cinlp-1.4
[14] A. C. Ribeiro, A. Sizo, H. L. Cardoso, and L. P. Reis, "Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis," in *EPIA 2021: Progress in Artificial Intelligence, Springer, , Online*, Cham: Springer International Publishing, 2021, pp. 766–777. https://doi.org/10.1007/978-3-030-86230-5_60
[15] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, "DeepSentipeer: Harnessing sentiment in review texts to recommend peer review decisions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019*, 2019, pp. 1120–1130. https://doi.org/10.18653/v1/P19-1106
[16] W. Jen and M. Chen, "Predicting Conference Paper Acceptance." 2018.

[17] A. Ghosh, N. Pande, R. Goel, R. Mujumdar, and S. S. Sistla, "Prediction, Conference Paper Acceptance (Acceptometer)." 2020.
[18] M. Skorikov and S. Momen, "Machine learning approach to predicting the acceptance of academic papers," in *The 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), IEEE*, 2020, pp. 113–117. https://doi.org/10.1109/IAICT50021.2020.9172011
[19] 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 *Proceedings of the First Workshop on Scholarly Document Processing, Online, 2020*, 2020, pp. 158–167. https://doi.org/10.18653/v1/2020.sdp-1.18
[20] A. Ciloglu and M. Merdan, "Big Peer Review Challenge." 2022.
[21] 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*, Singapore: Springer, 2021, pp. 143–152. https://doi.org/10.1007/978-981-33-4859-2_14
[22] 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, 2021. https://doi.org/10.1162/qss_a_00125
[23] P. Bao, W. Hong, and X. Li, "Predicting Paper Acceptance via Interpretable Decision Sets," in *The Web Conference 2021 - Companion of the World Wide Web Conference (WWW 2021), Ljubljana, Slovenia: Association for Computing Machinery (ACM)*, 2021, pp. 461–467. https://doi.org/10.1145/3442442.3451370
[24] P. K. Bharti, S. Ranjan, T. Ghosal, M. Agrawal, and A. Ekbal, "PEERAssist : Leveraging on Paper-Review Interactions to Predict Peer Review Decisions," in *International Conference on Asian Digital Libraries, Springer International Publishing, 2021*, Springer International Publishing, 2021, pp. 421–435. https://doi.org/10.1007/978-3-030-91669-5_33
[25] 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, 2021. https://doi.org/10.1016/j.neucom.2020.11.004
[26] 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. https://doi.org/10.1371/journal.pone.0010072
[27] S. Chakraborty, P. Goyal, and A. Mukherjee, "Aspect-based sentiment analysis of scientific reviews," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2020, pp. 207–216. https://doi.org/10.1145/3383583.3398541
[28] Z. J. Beasley, "Sentiment Analysis in Peer Review." 2020.
[29] M. Meng, R. Han, J. Zhong, H. Zhou, and C. Zhang, "Aspect-based sentiment analysis of online peer reviews and prediction of paper acceptance results," *DATA Sci. Inf.*, vol. 3, no. 1, 2023. https://doi.org/10.59494/dsi.2023.1.4
[30] S. Kumar, H. Arora, T. Ghosal, and A. Ekbal, "DeepASPeer: Towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Institute of Electrical and Electronics Engineers Inc., Jun. 2022*, Jun. 2022. https://doi.org/10.1145/3529372.3530937
[31] H. Arora, K. Shinde, and T. Ghosal, "Deciphering the Reviewer's Aspectual Perspective: A Joint Multitask Framework for Aspect and Sentiment Extraction from Scholarly Peer Reviews," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Institute of Electrical and Electronics Engineers Inc., 2023*, 2023, pp. 35–46. https://doi.org/10.1109/JCDL57899.2023.00015
[32] S. Basuki and M. Tsuchiya, "SDCF: semi-automatically structured dataset of citation functions," *Scientometrics*, vol. 127, no. 8, pp. 4569–4608, Aug. 2022. https://doi.org/10.1007/s11192-022-04471-x
[33] K. L. Lin and S. X. Sui, "Citation Functions in the Opening Phase of Research Articles: A Corpus-based Comparative Study," in *to Grammar, Media and Health Discourses - Part of The M. A. K. Halliday Library Functional Linguistics Series, no*, C. Approaches, Ed., Singapore: Springer, 2020, pp. 233–250. https://doi.org/10.1007/978-981-15-4771-3_10
[34] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles' metadata and cue-terms from content," *Scientometrics*, vol. 118, pp. 21–43, 2018. https://doi.org/10.1007/s11192-018-2961-x
[35] 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, pp. 1635–1684, 2019. https://doi.org/10.1007/s11192-019-03243-4
[36] S. Basuki and M. Tsuchiya, "The Quality Assist: A Technology-Assisted Peer Review Based on Citation Functions to Predict the Paper Quality," *IEEE Access*, vol. 10, pp. 126815–126831, 2022. https://doi.org/10.1109/ACCESS.2022.3225871
[37] W. Yuan, P. Liu, and G. Neubig, "Can We Automate Scientific Reviewing?," *J. Artif. Intell. Res.*, vol. 75, pp. 171–212, 2022. https://doi.org/10.1613/jair.1.12862
[38] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022. https://doi.org/10.1016/j.inffus.2021.11.011