EasyChair Preprint

№ 15080

Predicting Citation Counts with Machine Learning: a Citation Function Approach

Setio Basuki, Zamah Sari, Rizky Indrabayu, Reza Fauzan, Aulia Arif Wardhana and Masatoshi Tsuchiya

September 26, 2024

Predicting Citation Counts with Machine Learning: A Citation Function Approach

Setio Basuki
Informatics Engineering
Universitas Muhammadiyah Malang
Malang, Indonesia
setio_basuki@umm.ac.id

Zamah Sari
Informatics Engineering
Universitas Muhammadiyah Malang
Malang, Indonesia
zamahsari@umm.ac.id

Rizky Indrabayu
Informatics Engineering
Universitas Muhammadiyah Malang
Malang, Indonesia
bayyy17@webmail.umm.ac.id

Reza Fauzan
Department of Electical Engineering
Politeknik Negeri Banjarmasin
Banjarmasin, Indonesia
reza.fauzan@poliban.ac.id

Aulia Wardhana
Doctoral School
Wrocław University of Science and Technology
Wrocław, Poland
aulia.wardana@pwr.edu.pl

Masatoshi Tsuchiya
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
tsuchiya@is.cs.tut.ac.jp

***Abstract*— This paper develops a Machine Learning model to estimate the citation counts of research papers. The model uses citation functions, representing the intentions of the paper's author when making citations of previous works, to estimate the number of citations. These intentions include introducing a research topic, comparing, and criticizing previous works, etc. Three predictors have been developed based on citation functions: citing sentence, regular sentence, and reference. The prediction is seen as a regression and classification problem by pre-grouping the number of citations into high-count, medium-count, and low-count. The dataset was obtained from the International Conference on Learning Representations (ICLR) 2017-2020 containing 5,156 accepted and rejected papers. This paper uses only the accepted papers since the main task is to predict the number of citations of accepted/published papers. To obtain the number of citations one year after publication, this paper uses the API provided by Semantic Scholar. According to experiments, the best results in classification reach 98.33% accuracy, and in regression, the results reach 0.3 on both RMSE and MAE. The feature labeled 'citing paper dominant,' which reflects the superiority of the citing paper over the cited paper, has proven highly effective in delivering the best prediction results, even though it is sparsely represented in the dataset. In conclusion, citation function-based predictors are effective in estimating the future impact of a paper.**

Keywords—citation count, citation function, machine learning, number of citations, semantic scholar.

I. INTRODUCTION

The peer review process is a vital step in academic publishing and has become a standard practice in the scientific community for the publication of journals, conference papers, and grant proposals [1]. This process has become more challenging due to the massive number of paper submissions. The STM Report 2018 [2] states that there are more than 33,000 peer-reviewed journals (written in English) and over 9,000 journals (written in non-English) producing more than 3 million articles annually. A report by [3] states that it requires 15 million hours to review previously rejected manuscripts. The overload phenomenon also occurs in EasyChair, a conference management system that has handled more than 4 million users and around 100 thousand conference venues since 2002. More issues arise due to "reviewer fatigue" [4], [5]. This occurs because each submitted paper is assessed by two or more reviewers along with a handling editor. Additionally, potential reviewers want to review, but they receive too many review invitations [4]. This situation has put the peer review process into an overburdened system [6].

Citation functions represent the reasons behind in-text citations made by authors of research papers during the preparation of a research manuscript [7]. They come in many forms and represent different functions [8], such as introducing the research topic, showing research trends, comparing and contrasting, and extending previous works. Citations provide many benefits when assessing a paper's quality during the review process. They help clarify the research's position within the broader literature [9], offer a clear view of the paper's main topic [10], emphasize its novelty and originality [11], and aid in evaluating the overall quality of the manuscript [12], [13]. Given the important roles of citation functions, they could potentially serve as predictors for estimating a paper's future impact.

Even though research on estimating future citation counts has gained much attention, no single study uses the reason behind citations (citation function) as the main predictor, considering its important position in the paper. Most existing research estimates citation counts automatically, as in [14], [15], [16], [17], [18], while other research uses similar concepts with different terminology, such as predicting paper popularity [19], predicting highly cited academic paper [20], predicting the future impact of publications [21], and top paper prediction [22]. **Therefore, this paper aims to predict the citation counts obtained by research papers one year after publication**. Research by [23] stated that the accumulation of citations one and two years after publication might serve as a forward indicator of the long-term quality of research publications. This finding is supported by [24], which suggests that citations received in the first year contribute to the accurate prediction of long-term citation impact. The prediction consists of two ML tasks: (1) classification of three pre-grouped citation counts (high-count, medium-count, low-count) and (2) regression to directly predict the number of citations. The ML models for prediction are constructed based

on citation functions, which represent the intentions of paper authors when citing previous related works. The model will be realized through citation function-based predictors: (1) citing sentence predictor, representing sentences in the paper containing citation marks; (2) regular sentence predictor refers to sentences within the paper that do not include citation marks; and (3) one additional predictor called the reference-based predictor, which captures the role of reference types in making predictions. For experimental purposes, this paper incorporates these predictors into a combination predictor to predict citation counts. The models are trained using the paper repository from the International Conference on Learning Representations (ICLR) 2017-2020, containing 5,156 accepted and rejected papers. The classification task utilizes the Extreme Gradient Boosting (XGBoost) algorithm, assessed by accuracy, while the regression task employs the Extreme Gradient Boosting for Regression (XGBR) algorithm, evaluated with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Moreover, model development for classification and regression implements a feature selection method called Chi-Square (Chi2).

This paper delivers several contributions. Firstly, it uses citation functions to develop Machine Learning models for predicting citation counts. Secondly, it employs three predictors—citing sentence, regular sentence, and reference—to enhance the accuracy of citation count predictions. Thirdly, the models achieve competitive classification results with 98.33% accuracy in classifying high, medium, and low-count papers and significant regression results with an RMSE and MAE of 0.3 when predicting the number of citations for high-count papers. Additionally, an analysis of the top 10 most influential features indicates that 'citing paper dominant,' which reflects the superiority of the citing paper over the cited paper, has proven highly effective in delivering the best prediction results. Overall, citation functions have a strong relationship with the future impact of research papers.

II. METHOD FOR CITATION COUNT PREDICTION

This section shows how the proposed prediction system for estimating the citation count of scientific papers is developed. The prediction system is developed using citation functions that illustrate the reasons authors cite earlier works in their academic papers. This section will cover several key aspects: (a) the ICLR paper dataset and the sources of citation counts for each paper, (b) the prediction features, which include citing sentence features, regular sentence features, and reference-based features, and (c) the prediction scenario. It is important to note that the prediction of citation counts is treated as a regression problem.

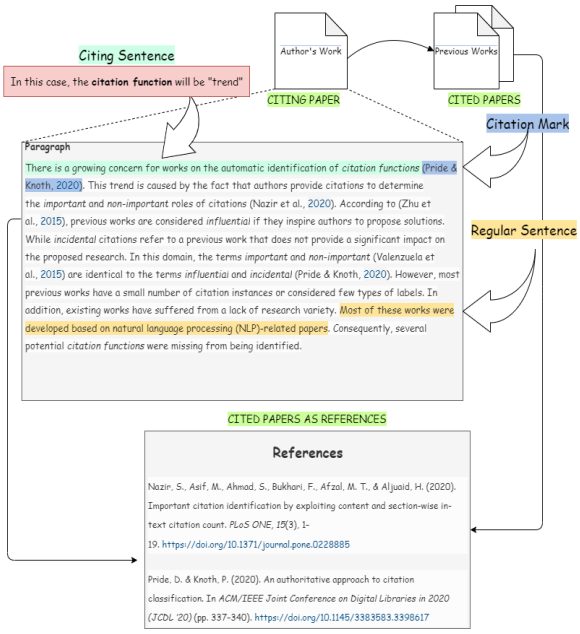

Fig. 1. The position of citing paper, cited paper, citing sentence, and regular sentence in the paper.

There are important technical terms used in this paper. A citing paper refers to a paper that references other or earlier papers, while a cited paper is one that is referenced by a citing paper. A citing sentence includes a citation mark, whereas regular sentences do not contain any citation marks. Figure 1 provides an illustration of these terms.

A. *Research Paper Dataset*

The dataset of papers was sourced from the International Conference on Learning Representations (ICLR) from 2017 to 2020 and includes 5,156 papers [25], as presented in Table I. The acceptance or rejection of each paper by the conference editor is documented, while the quality indicator (good or poor) is determined by the average review score according to the research [26]. The number of citations each paper received was automatically collected from Semantic Scholar one year after its publication date. For example, the number of citations for a paper published in 2017 would be collected in 2018. This approach allows the citation counts to be compared among papers published in different years.

TABLE I. THE ICLR PAPERS 2017-2018 USED AS DATASET.

Publication Period	Accepted	Rejected	Good	Poor	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**

However, not all papers in the dataset can be used to build the prediction features because the prediction will only be made on accepted/published papers. Thus, the prediction dataset consists of 1,722 paper instances as shown in the "Accepted Column" of Table I.

B. Citing Sentence Features

The main feature used in this paper is citing sentences, which represent sentences in the research paper containing citation marks. This feature is generated by extracting all citing sentences from the papers in the dataset and categorizing them into 18 primary labels (1 to 18) of citation functions, as outlined in Table II. Additionally, we added 2 more features: the "Other" label (number 19) to accommodate citing sentences that cannot be categorized into the main labels, and the last label (number 20) constructed by calculating the presence of each feature's label in each paper. The final citing sentence label consists of 20 attributes. The model development through classification in this stage is performed using the Bidirectional Encoder Representations from Transformers (BERT)-based model developed by [27].

TABLE II. THE CITING SENTENCE FEATURES.

General (Coarse) Label "Background"
Background information, such as theory, principle, topic, etc.
Detailed Label (Fine-grained): 1. **definition:** definition of concept, theory, topic, or problem. *Example:* U-tree <citation> is an online agent algorithm designed to identify a compact state representation from a continuous stream of experiences. **2. suggest,** encouraging the reader to explore the cited papers in more detail. *Example*: For more detailed information, we refer the reader to <citation> or <citation>. 3. **judgment,** highlighting the positive and negative aspects, usefulness or limitations, and other attributes of a concept, topic, etc. *Example*: Secondly, it can be argued that a measure like F1 is inappropriate for chunking tasks <citation>. 4. **technical,** explaining how a theory, principle, concept, topic, or problem works. *Example*: Traditionally, motion fields are estimated using the variational model proposed by Horn and Schunck <citation>. 5. **trend,** describing the importance of the research topic, theory, principle, concept, or problem. *Example*: One of the most widely used algorithms for blind source separation (BS) is the FastICA (FICA) algorithm <citation>.
General (Coarse) Label "Citing Paper Work"
The work/research that is proposed by the author
Detailed Label (Fine-grained): 6. **corroboration,** when proposing a research topic, the citing paper references the cited paper. *Example*: To accomplish this, we build on the idea of continuous regression <citation>. 7. **based on,** indicating that the citing paper follows, considers, is based on, and is inspired by the cited paper. *Example*: For a thorough overview of network coding theory, please see <citation>. 8. **use,** the citing paper utilizes, implements, employs, or adopts the concept, dataset, or technique. *example*: We use an algorithm derived from the 0-efficiency technique developed by Jaco and Rubinstein <citation>. 9. **extend,** the citing paper enhances, supplements, or alters the work presented in the cited paper. *Example*: We further refine the upper bound for the general outerplanar graph from the <formula> given in <citation>. 10. **dominant,** the citing paper demonstrates better performance than the cited paper. *Example*: When <formula>, our method surpasses BM3D by 0.7 dB, achieving the predicted upper bound over BM3D as noted in <citation>. 11. **future,** outlining the future directions of the citing paper. *Example*: In future research, we plan to incorporate the concept from <citation> into our watermarking algorithm.
General (Coarse) Label "Cited Paper Work"
The work/research has been done by cited papers (previous work).
Detailed Label (Fine-grained): 12. **propose,** explaining the research proposed by the cited paper. *Example*: Another method <citation> aims to minimize an energy functional and obtain the most likely segmentation from a global perspective. 13. **success,** emphasizing the success of cited paper. *Example*: Larose and Tesson <citation> effectively applied the theory to explore finer complexity classes of CSPs. 14. **weakness,** pointing the limitation/drawback of cited paper. *Example*: The Viola and Jones algorithm <citation> did not successfully detect faces in a large number of frames. 15. **result,** explaining the results of the cited paper neutrally. *Example*: The theorem provides a sufficient condition for a broader class of operators, and it generalizes the result in <citation>. 16. **dominant,** noting the superiority of the cited paper over the citing paper. *Example*: Chan et al. <citation> propose a probabilistic approach to achieve a lower number of tests, which outperforms our scheme.
General (Coarse) Label "Compare and Contrast"
Analyzing the similarities and differences between citing and cited papers.
Detailed Label (Fine-grained): 17. **compare,** explaining the similarity between citing and cited papers. *Example*: Methods for dynamic word embeddings <citation> are closely aligned with our research focus. 18. **contrast,** explaining the differences between citing and cited papers. *Example*: It is noteworthy that unlike <citation>, we retrain both pruned networks only once.
General (Coarse) Label "Other"
Accommodating for citing sentences that do not align with any of the above indicators.
Detailed Label (Fine-grained): 19. **comparison,** a comparison of the cited papers (whether they are similar or different is unclear). *Example*: These methods include the support vector metric learning algorithm developed by Xu et al. <citation>, the gradient-boosted large margin nearest neighbor approach introduced by Kedem et al. <citation>, and the Hamming distance metric learning technique designed by Norouzi et al. <citation>. 19. **multiple_intent,** citing sentences contain two or more citation marks, each serving a different purpose. *Example*: The table compares the computational complexity of the proposed method against AOG <citation> and NCTE <citation>. 19. **other,** this label is intended for citing sentences that do not fit into any of the categories described above. *Example*: Between them, Kikuchi's cluster variational method <citation>.

C. Regular Sentence Features

Applying the citation function labeling scheme used for citing sentences, we categorize regular sentences with the BERT model into 18 labels (see Table II). We also add two extra labels: "Other" and a label showing the presence of each category in the dataset's papers.

D. Reference-based Features

The reference-based features are additional features developed to clarify the impact of the source of citations (references) in the prediction process. This type of feature consists of 24 labels which can be divided into several categories: generic, preprint, conference, and journal. To develop this feature, this paper extracts all reference sections from each paper in the dataset. Following this, we calculate the presence of each label using a rule-based approach. The detailed reference-based features are shown in Table III.

TABLE III. THE LIST OF FEATURES REPRESENTING REFERENCES

Generic Labels
1. Number of total references 2. Count of references from the last 3 years
Pre-Print Labels
3. Preprint Repository (arXiv)
Conference Labels

4.	Conference on Neural Information Processing Systems
5.	International Conference on Learning Representations
6.	International Conference on Machine Learning
7.	Association for the Advancement of Artificial Intelligence
8.	International Conference on Computer Vision
9.	Conference on Computer Vision and Pattern Recognition
10.	Empirical Methods in Natural Language Processing
11.	Association for Computational Linguistics
12.	North American Chapter of the Association for Computational Linguistics
13.	European Conference on Computer Vision
14.	The International Conference on Robotics and Automation
15.	the International Conference on Acoustics, Speech, and Signal Processing
16.	The International Joint Conference on Artificial Intelligence
17.	The International Conference on Artificial Intelligence and Statistics
18.	Special Interest Group on Knowledge Discovery and Data Mining
Journal Labels	
19.	Neural Computation
20.	**IEEE Transaction**
21.	**ACM Transaction**
22.	**MIT Press**
23.	**Nature**
24.	**JMLR**: The Journal of Machine Learning Research

E. The experiment scenario of citation count prediction

The experiment is designed as follows. The citation count prediction is seen as a regression problem with the number of obtained citations one year after publication as the target prediction. There are four predictors: citing sentence, regular sentence, reference, and combination (64 features are a combination of citing sentence, regular sentence, and reference-based predictors). The regression algorithm used in this paper is Extreme Gradient Boosting Regression (XGBR), combined with Chi-Square (Chi2) as a feature selection technique, to develop the regression models. This paper proposes several data preprocessing and regression stages. First, data normalization (scale 1-10) on the target prediction attribute (citation counts) is performed to handle the wide values gap among the papers (several papers obtained no citations, while others received thousands of citations). Second, data normalization is applied to all attributes. Third, papers are grouped into high-count (>=7), medium-count (4-6), and low-count (1-3) categories according to the normalized target attribute. Fourth, both the classifications and regressions are implemented using each predictor, i.e., citing sentence, regular sentence, reference, and combination. For classification, this paper implements oversampling to make the dataset (training data only) more balanced. Oversampling works by randomly duplicating instances from the minority class until the dataset reaches the desired ratio while keeping the number of instances in the majority class unchanged. For regression, each prediction is applied to each group. Finally, while classification is measured based on accuracy, regression is measured using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

III. Experiment Results and Discussion

This section demonstrates the experimental results of the prediction. The results can be divided into two parts: prediction performance in classification and regression, and analysis of the most influential features.

A. Prediction Performance

Table IV shows the best regression results for all citation categories. The model performs well in predicting citations for high-count papers, with RMSE and MAE both at 0.3 using the citing sentence predictor. Lower scores for RMSE and MAE indicate better performance. The models are still quite effective in estimating the number of citations obtained by low-count papers (0.6 RMSE and 0.5 MAE) using the regular sentence predictor and medium-count papers (0.7 RMSE and 0.5 MAE) using the reference and combination-based predictor. The next **notable finding** is that the citation function-based predictors, citing sentences and regular sentences, prove their effectiveness in achieving the best results.

This paper has also conducted experiments on classifying three groups of citation counts, as shown in Table V. The classifications are implemented using two types of datasets: the original and the oversampled. The dataset needs to be oversampled because of the imbalanced distribution among the three categories. Generally, there is an improvement of more than 5% in accuracy when using the oversampled dataset. It is observed that the highest accuracy is achieved by using the combination predictor, showing 98.33% on the oversampled dataset, using almost all features (55 features). Interestingly, an almost similar accuracy of 98.22% is reached by the regular sentence predictor, using only 18 features. The overall results in the classification setting show the effectiveness of citation function-based predictors in understanding the pattern of the three different categories. By using fewer features in the original dataset setting, the accuracies of around 92% are identical to the other predictors (reference and combination). A similar trend is also observed in the oversampled setting, where the citation function-based predictor reaches competitive accuracies compared to the combination predictor.

The competitive performances in both regression and classification strengthen our hypothesis that the authors' intentions when making citations are closely associated with the quality of research papers, especially in predicting the future popularity of the research paper.

B. Feature Analysis of Predictors

The next discussion is the analysis of the top 10 most influential features for prediction to obtain the best performance in regression and classification tasks.

On the regression task, the analysis of the top 10 most important features is conducted for each category. In the low-count category, the feature 'citing paper dominant' appears four times in the citing sentence, regular sentence, and combination predictors. In the medium-count category, the features are dominated by 'citing paper based on,' 'citing paper corroboration,' and 'citing paper extend.' As in the low-count category, the features representing performance dominance, i.e., 'citing paper dominant' and 'cited paper dominant,' appear four times in the high-count category. These types of features have low distribution (please refer to our previous publications [26], [27]) but have a significant impact on prediction. The selected features are shown in Table VI.

On the classification task, the citing sentence, reference, and combination predictors show the same top 10 most influential features in both the original and resampled datasets. Six features belonging to citation function-based predictors dominate the prediction, namely: cited paper success, cited

paper weaknesses, citing paper use, citing paper future, citing paper dominance, and technical. This dominance is supported by their presence (except for 'technical') in the combination predictor. Interestingly, the feature 'citing paper dominant' appears in all scenarios, although its distribution in the whole dataset is relatively low. This feature aligns with the common practice in writing scientific papers in the computer science domain, demonstrating that the performance of the citing paper outperforms previous related works. The details of the most influential features are presented in Table VII.

TABLE IV. THE BEST REGRESSION PERFORMANCE ON EACH PREDICTOR.

Predictors	**LOW-COUNT**			**MEDIUM-COUNT**			**HIGH-COUNT**		
	N	**MAE**	**RMSE**	**N**	**MAE**	**RMSE**	**N**	**MAE**	**RMSE**
citing sentence	1	0.514899	0.652723	1	0.667430	0.886717	12	**0.322970**	**0.379312**
regular sentence	20	0.535915	0.697645	10	**0.570814**	**0.763458**	3	0.729759	0.931158
reference-based	1	0.506379	**0.646072**	1	0.670931	0.792426	20	0.823868	1.055666
combination	48	**0.502040**	0.664193	16	0.603760	0.792928	11	0.458484	0.522735

TABLE V. THE BEST CLASSIFICATION RESULTS FOR HIGH, MEDIUM, AND LOW-COUNT GROUPS

Predictors	**Original Data**		**Oversampled Data**	
	N	**Accuracy (%)**	**N**	**Accuracy (%)**
citing sentence	16	92.33	13	97.55
regular sentence	1	92.33	18	98.22
reference-based	23	92.64	22	97.78
combination	13	92.64	55	**98.33**

TABLE VI. TOP 10 MOST IMPACTFUL FEATURES OF EACH PREDICTOR FOR OPTIMAL REGRESSION PERFORMANCE

LOW-COUNT				
Rank	**#1 predictor: Citing sentence**	**#2 predictor: Regular sentence**	**#3 predictor: Reference**	**#4 predictor: Combination**
1.	#1-citing paper: dominant	#2-other	#3-num_ref_3years	#3-num_ref_3years
2.		#2-citing paper: dominant		#1-citing paper: dominant
3.		#2-citing paper: extend		#1-citing paper: based on
4.		#2-definition		#3-arxiv
5.		#2-judgment		#1-num. of citing sentence
6.		#2-compare		#2-other
7.		#2-cited paper: weakness		#2-citing paper: dominant
8.		#2-citing paper: future		#3-naacl
9.		#2-citing paper: use		#3-neurips
10.		#2-cited paper: dominant		#3-emnlp
MEDIUM-COUNT				
Rank	**#1 predictor: Citing sentence**	**#2 predictor: Regular sentence**	**#3 predictor: Reference**	**#4 predictor: Combination**
1.	#1-compare	#2-compare	#3-cvpr	#1-compare
2.		#2-citing paper: based on		#3-cvpr
3.		#2-citing paper: extend		#2-compare
4.		#2-citing paper: corroboration		#1-trend
5.		#2-num. of regular sentence		#2-citing paper: based on
6.		#2-other		#1-citing paper: corroboration
7.		#2-judgment		#2-citing paper: extend
8.		#2-citing paper: future		#2-citing paper: corroboration
9.		#2-cited paper: success		#3-neuralcom
10.		#2-citing paper: use		#3-iccv
HIGH-COUNT				
Rank	**#1 predictor: Citing sentence**	**#2 predictor: Regular sentence**	**#3 predictor: Reference**	**# predictor: Combination**
1.	#1-other	#2-num. of regular sentence	#3-jmlr	#2-num. of regular sentence
2.	#1-cited paper: dominant	#2-other	#3-aistats	#1-other
3.	#1-citing paper: dominant	#2-judgment	#3-icassp	#3-jmlr
4.	#1-definition		#3-iccv	#1-cited paper: dominant
5.	#1-citing paper: extend		#3-ijcai	#3-aistats
6.	#1-compare		#3-num_ref_3years	#2-other
7.	#1-num. of citing sentence		#3-neuralcom	#1-citing paper: dominant
8.	#1-citing paper: corroboration		#3-ieee_tran	#2-judgment
9.	#1-cited paper: propose		#3-mit_press	#1-definition
10.	#1-suggest		#3-acm_tran	#3-icassp

TABLE VII. TOP 10 MOST IMPACTFUL FEATURES OF EACH PREDICTOR FOR ACHIEVING THE HIGHEST CLASSIFICATION PERFORMANCE

Original Dataset				
Rank	**#1 predictor: Citing sentence**	**#2 predictor: Regular sentence**	**#3 predictor: Reference**	**#4 predictor: Combination**
1.	#1-cited paper: success	#2-other	#3-num_ref_3years	#3-num_ref_3years
2.	#1-cited paper: weaknesses		#3-arxiv	#3-arxiv
3.	#1-citing paper: use		#3-aistats	#1-cited paper: success
4.	#1-citing paper: future		#3-acl	#2-other
5.	#1-citing paper: dominant		#3-naacl	#1-cited paper: weaknesses
6.	#1-citing paper: extend		#3-emnlp	#3-aistats
7.	#1-trend		#3-icml	#2-cited paper: success
8.	#1-technical		#3-eccv	#1-citing paper: use
9.	#1-cited paper: dominant		#3-cvpr	#1-citing paper: future
10.	#1-definition		#3-mit_press	#3-acl
Oversampled Dataset				
Rank	**#1 predictor: Citing sentence**	**#2 predictor: Regular sentence**	**#3 predictor: Reference**	**#4 predictor: Combination**
1.	#1-cited paper: success	#2-other	#3-num_ref_3years	#3-num_ref_3years
2.	#1-cited paper: weaknesses	#2-cited paper: success	#3-arxiv	#3-arxiv
3.	#1-citing paper: use	#2-num. of regular sentence	#3-aistats	#1-cited paper: success
4.	#1-citing paper: future	#2-cited paper: propose	#3-acl	#2-other
5.	#1-citing paper: dominant	#2-citing paper: corroboration	#3-naacl	#1-cited paper: weaknesses
6.	#1-citing paper: extend	#2-cited paper: weakness	#3-emnlp	#3-aistats
7.	#1-trend	#2-citing paper: dominant	#3-icml	#2-cited paper: success
8.	#1-technical	#2-technical	#3-eccv	#1-citing paper: use
9.	#1-cited paper: dominant	#2-citing paper: future	#3-cvpr	#1-citing paper: future
10.	#1-definition	#2-citing paper: use	#3-mit_press	#3-acl

IV. CONCLUSION

This paper presents a Machine Learning model designed to predict citation counts for research papers. The model leverages citation functions to understand the reasons for citations by authors. Predictions are approached as both classification problems (categorizing papers into high, medium, or low citation counts) and regression problems (predicting the exact number of citations). Our experiments show that the models were effective in both tasks, demonstrating classification accuracies of around 98% and competitive regression results with RMSE and MAE of 0.3. An RMSE of 0.3 is a competitive result because it is achieved within a 3-point interval (low: 1 to 3, medium: 4 to 6, and high: >= 7). This means that the RMSE is around 10% of the interval range. For more comprehensive results, several potential methods can be proposed, including analyzing whether the predictors can be used to detect "sleeping beauty" papers (papers that remain unnoticed for years before obtaining significant citations as they are acknowledged for their importance).

The prediction performances indicate that the citation functions are strongly correlated with the quality of the papers, particularly in estimating future paper impact. Additionally, we observe that the feature 'citing paper dominant' which reflects the superior performance of the citing paper over the cited paper, has proven highly effective in delivering the best results in all scenarios, even though it is infrequently represented in the dataset. This phenomenon aligns with the common practice in the computer science domain, where authors often claim the superiority of their works compared to previous related works.

REFERENCES

[1] F. Rowland, "The peer-review," *Learned Publishing*, vol. 15, no. 4, pp. 247–258, 2002, doi: https://doi.org/10.1087/095315102760319206.

[2] R. Johnson, A. Watkinson, and M. Mabe, "The STM Report - An overview of scientific and scholarly publishing," The Hague, The Netherlands, Oct. 2018. [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," *Humanit Soc Sci Commun*, vol. 8, no. 1, Dec. 2021, doi: 10.1057/s41599-020-00703-8.

[4] M. Breuning, J. Backstrom, J. Brannon, B. I. Gross, and M. Widmeier, "Reviewer Fatigue? Why Scholars Decline to Review their Peers' Work," *PS Polit Sci Polit*, vol. 48, no. 4, pp. 595–600, 2015, doi: 10.1017/S1049096515000827.

[5] C. W. Fox, A. Y. K. Albert, and T. H. Vines, "Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution," *Res Integr Peer Rev*, vol. 2, no. 1, pp. 1–6, 2017, doi: 10.1186/s41073-017-0027-x.

[6] M. Jubb, "Peer review: The current landscape and future trends," *Learned Publishing*, vol. 29, no. 1, pp. 13–21, 2016, doi: 10.1002/leap.1008.

[7] S. Teufel, A. Siddharthan, and D. Tidhar, "An annotation scheme for citation function," in *COLING/ACL 2006 - SIGdial06: 7th SIGdial Workshop on Discourse and Dialogue, Proceedings of the Workshop*, 2006, pp. 80–87. doi: 10.3115/1654595.1654612.

[8] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2015.

[9] K. L. Lin and S. X. Sui, "Citation Functions in the Opening Phase of Research Articles: A Corpus-based Comparative Study," *Corpus-based Approaches to Grammar, Media and Health Discourses - Part of The M.A.K. Halliday Library Functional Linguistics Series*, no. Springer Singapore, pp. 233–250, 2020, doi: https://doi.org/10.1007/978-981-15-4771-3_10.

[10] 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, doi: 10.1007/s11192-018-2961-x.

[11] 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, doi: 10.1007/s11192-019-03243-4.

[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. [Online]. Available: http://ceur-ws.org/Vol-2414/

[13] A. S. Raamkumar, S. Foo, and N. Pang, "What papers should I cite from my reading list? User evaluation of a Manuscript Preparatory assistive task," *CEUR Workshop Proc*, vol. 1610, no. i, pp. 51–62, 2016.

[14] J. Hirako, R. Sasano, and K. Takeda, "Realistic Citation Count Prediction Task for Newly Published Papers," in *Findings ofthe Association for Computational Linguistics: EACL*, May 2023, pp. 1131–1141. [Online]. Available: https://arxiv.org/list/cs.CL/recent

[15] Y. A. Alohali, M. S. Fayed, T. Mesallam, Y. Abdelsamad, F. Almuhawas, and A. Hagr, "A Machine Learning Model to Predict Citation Counts of Scientific Papers in Otology Field," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/2239152.

[16] A. Pandey Akella, H. Alhoori, P. Ravikanth Kondamudi, C. Freeman, and H. Zhou, "Early Indicators of Scientific Impact: Predicting Citations with Altmetrics," *J Informetr*, vol. 15, no. 2, 2021, [Online]. Available: https://www.elsevier.com/open-access/userlicense/1.0/

[17] K. Kousha and M. Thelwall, "Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper," Mar. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1002/asi.24810.

[18] X. Ruan, Y. Zhu, J. Li, and Y. Cheng, "Predicting the citation counts of individual papers via a BP neural network," *J Informetr*, vol. 14, no. 3, Aug. 2020, doi: 10.1016/j.joi.2020.101039.

[19] K. Abbas *et al.*, "Predicting the Future Popularity of Academic Publications Using Deep Learning by Considering It as Temporal Citation Networks," *IEEE Access*, vol. 11, pp. 83052–83068, 2023, doi: 10.1109/ACCESS.2023.3290906.

[20] X. Tang, H. Zhou, and S. Li, "Predictable by publication: discovery of early highly cited academic papers based on their own features," *Library Hi Tech*, vol. ahead-of-print, no. ahead-of-print, Jan. 2023, doi: 10.1108/LHT-06-2022-0305.

[21] S. Huang, Y. Huang, Y. Bu, Z. Luo, and W. Lu, "Disclosing the interactive mechanism behind scientists' topic selection behavior from the perspective of the productivity and the impact," *J Informetr*, vol. 17, no. 2, p. 101409, 2023, doi: https://doi.org/10.1016/j.joi.2023.101409.

[22] Z. Sun, "Textual features of peer review predict top-cited papers: An interpretable machine learning perspective," *J Informetr*, vol. 18, no. 2, p. 101501, 2024, doi: https://doi.org/10.1016/j.joi.2024.101501.

[23] J. Adams, "Early citation counts correlate with accumulated impact," *Scientometrics*, vol. 63, no. 3, pp. 567–581, 2005, doi: 10.1007/s11192-005-0228-9.

[24] C. Stegehuis, N. Litvak, and L. Waltman, "Predicting the long-term citation impact of recent publications," *J Informetr*, vol. 9, no. 3, pp. 642–657, 2015, doi: https://doi.org/10.1016/j.joi.2015.06.005.

[25] W. Yuan, P. Liu, and G. Neubig, "Can We Automate Scientific Reviewing?," *Journal of Artificial Intelligence Research*, vol. 75, pp. 171–212, 2022, doi: https://doi.org/10.1613/jair.1.12862.

[26] 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, doi: 10.1109/ACCESS.2022.3225871.

[27] 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.