

# Classification of the Top-cited Literature by Fusing Linguistic and Citation Information with the Transformer Model

Masanao Ochi<sup>1</sup> <sup>a</sup>, Masanori Shiro<sup>2</sup>, Jun'ichiro Mori<sup>1</sup> and Ichiro Sakata<sup>1</sup>

<sup>1</sup>*Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan*

<sup>2</sup>*HIRI, National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba, Ibaraki, Japan*

**Keywords:** Citation Analysis, Scientific Impact, Graph Neural Network, BERT.

**Abstract:** The scientific literature contains a wide variety of data, including language, citations, and images of figures and tables. The Transformer model, released in 2017, was initially used in natural language processing but has since been widely used in various fields, including image processing and network science. Many Transformer models trained with an extensive data set are available, and we can apply small new data to the models for our focused tasks. However, classification and regression studies for scholarly data have been conducted primarily by using each data set individually and combining the extracted features, with insufficient consideration given to the interactions among the data. In this paper, we propose an end2end fusion method for linguistic and citation information in scholarly literature data using the Transformer model. The proposed method shows the potential to efficiently improve the accuracy of various classifications and predictions by end2end fusion of various data in the scholarly literature. Using a dataset from the Web of Science, we classified papers with the top 20% citation counts three years after publication. The results show that the proposed method improves the F-value by 2.65 to 6.08 percentage points compared to using only particular information.

## 1 INTRODUCTION

The early detection of promising research is vital to identify research worthy of investment. Additionally, to the increasing number of digital publications in the scholarly literature and the fragmentation of research, there is a need to develop techniques to predict future research trends automatically. Previous studies on impact prediction of scholarly literature have used features specifically designed for each indicator (Ayaz et al., 2018; Miró et al., 2017; Schreiber, 2013; Acuna et al., 2012; Bai et al., 2019; Sasaki et al., 2016; Stegehuis et al., 2015; Cao et al., 2016) or a link prediction using custom networks (Yan and Guns, 2014; Park and Yoon, 2018; Yi et al., 2018).

However, recent advances in deep learning technology have facilitated integrating different individual models and constructing more general-purpose models, such as the Transformer model (Vaswani et al., 2017). The Transformer model, released in 2017, was initially used in natural language processing (Devlin et al., 2019) but has since been widely used in various fields, including image processing (Dosovitskiy

et al., 2021) and network science (Zhang et al., 2020). This model has several advantages, including publishing trained models with large datasets and fine-tuning by applying new data to individual tasks.

Against this backdrop, the impact of the academic literature has been evaluated either by creating individual features or by formulating the problem as a network link prediction problem. However, the rise of general-purpose models such as the Transformer can transform this situation.

Scholarly literature contains various data, including language, citations, and images of figures and tables. Several studies have pointed out that network information, rather than linguistic information, may be necessary for predicting the impact of scholarly literature (Sasaki et al., 2016; Ochi et al., 2021). In particular, Ochi et al. report that citation networks may be more biased than linguistic information in the embedding space of papers with future high citations (Ochi et al., 2021). This result indicates the need to develop a more advanced model than the BERT model using only linguistic information in the academic literature, such as the SPECTOR model (Cohan et al., 2020), with the top-cited papers as teacher data.

<sup>a</sup>  <https://orcid.org/0000-0002-6661-6735>

In this paper, we propose an end2end fusion method of linguistic and citation information in scholarly literature data using the Transformer model. Using a dataset extracted from the Web of Science, we evaluated the proposed method for classifying papers with the top 20% of citations three years after publication. We found that the proposed method improved the F-value by 2.65 to 6.08 points compared to using only individual information. This method makes it possible to fuse diverse data from the scholarly literature into end2end. Experimental results also show the possibility of efficiently improving the accuracy of various classifications and predictions. Our proposed method is threefold.

- We developed an end2end model that fuses linguistic features and a citation network of scholarly literature data.
- The proposed model automatically selects when citation network information is valid and when linguistic information is valid.
- The proposed model improves the classification accuracy of the papers with the highest number of citations after five years.

The remainder of this article will first introduce the related work in Section 2. At this point, we describe the context of prediction of scholarly impact and clarify the needs of the end2end model, which fuses linguistic and citation information. Section 3 describes our proposed model, including its architecture. Section 4 reveals the experiment in detail. A discussion of the results appears in Section 5. Finally, Section 6 emphasizes the scientific contribution of the work and notes several challenges we can address in the future.

## 2 RELATED WORK

In this section, we contextualize the Transformer model, describe its application and extension to scholarly literature data, and then describe the research conducted on the index, the influence of scholarly literature, its predictions, and challenges, and clarify the position of this study.

### 2.1 Transformer Model for Scholarly Data

The Transformer model(Vaswani et al., 2017), one of the Encoder-Decoder models using Attention, is capable of large-scale learning due to its slight computational complexity and parallel computing capability. The Transformer model was quickly put to use

when the BERT model(Devlin et al., 2019) showed the highest accuracy on the GLUE dataset(Wang et al., 2018), a multi-task accuracy competition for natural language processing. Since then, its use has expanded in diverse fields, such as image processing(Dosovitskiy et al., 2021) and network science(Zhang et al., 2020).

The application of the Transformer model to scholarly literature data is also underway. The first is the SciBERT model(Beltagy et al., 2019), which is based on the BERT model and trained on text data from academic literature. SciBERT focuses on generic Embedding acquisition for academic literature at the word level. However, the SPECTER model(Cohan et al., 2020) attempts to obtain Embedding at the paper level rather than at the word level. The SPECTER model acquires Embedding at the paper level by making the papers that have a citation relationship with each other a pair of positive examples.

### 2.2 The Influence of Scholarly Literature

However, scholarly literature contains not only text but also various types of information such as citations, figures, tables, authors, and institutional affiliations. Researchers used this information to index the influence of academic literature, for example, the number of citations, *h*-index for authors(Hirsch, 2005), Journal Impact Factor (JIF) for journals(Garfield and Sher, 1963), and Nature Index (NI) for research institutions. Many studies have predicted future *h*-index values (Ayaz et al., 2018; Miró et al., 2017; Schreiber, 2013; Acuna et al., 2012). Acuna et al. calculated an equation for predicting the *h*-index. They showed that five main parameters are fundamentally crucial for prediction (Acuna et al., 2012): the number of publications, the current *h*-index value, the number of years since the first publication, the number of types of journals published to date, and the number of papers in top journals.

There are some studies to predict the number of future citations of papers(Bai et al., 2019; Sasaki et al., 2016; Stegehuis et al., 2015; Cao et al., 2016). Among them, Stegehuis et al. and Cao et al. predict the number of citations in the far future, considering the number of citations during 1–3 years after publication. In contrast, Sasaki et al. predict the number of citations after three years from publication directly(Sasaki et al., 2016). The task evaluated in this study also predicts the number of citations three years after publication, just as Sasaki et al. did. Previous efforts to predict indicators have created various fea-

tures and used them as input to the model.

There are attempts to predict the impact of scholarly literature more directly as a link prediction problem by creating a custom network. Yan et al. evaluated the impact of academic literature by creating a co-author network of countries, institutions, and authors and predicting their link relationships (Yan and Guns, 2014). They showed that predicting author coauthorship was more difficult than predicting country or institution coauthorship. Park et al. created a citation network of patent information between the two fields and developed a model to predict future trends in the number of citations across fields (Park and Yoon, 2018). They used it to predict increasing trends in linkages between the biotechnology field and the information technology field, showing that technological convergence is underway. Yi et al. constructed a bipartite graph, author, and keywords from the scholarly literature data (Yi et al., 2018). With this, they developed a model to predict future changes in author interest. By evaluating the model as a link prediction problem between authors and keywords, they show that it can predict future changes in each author's interest based on past trends in authors' keywords. Thus, the direct use of network information effectively predicts the influence of academic literature.

However, studies that used each data separately or combined the extracted features for classification or regression did not adequately consider the interactions among the data. It is also a challenge to make a more active use of citation information rather than simply using it as teacher data, as in the SPECTER model. In particular, it is vital to build end2end models that fuse various academic literature data to build more general-purpose models. As a first step, this paper proposes an end2end fusion method of linguistic and citation information in academic literature data using the Transformer model.

### 3 PROPOSED METHOD

We plan to build a model that can learn end2end by fusing linguistic and citation information among the diverse data possessed by the academic literature. However, is it necessary to fuse multiple pieces of information to predict the impact of academic literature? If a model can fully understand the text of a paper, is it sufficient to predict the impact of that paper? This section shows that citation information may be more important than a paper's content in predicting its impact. That is, we require a model that actively incorporates citation information. We propose

a model that can be trained end2end by fusing linguistic and citation information.

#### 3.1 Linguistic or Citation Information?

Is it necessary to fuse language and citation information to predict the impact of academic literature? Is it impossible to predict the scholarly literature's impact if the model accurately understands the linguistic content? Several studies have provided rebuttal evidence to this question. Sasaki et al. constructed a linear model to predict the number of citations and reported that the features associated with the citation network are important (Sasaki et al., 2016). Ochi et al. used a network embedding and a language model to examine how methods to place the top-cited papers in the embedding space (Ochi et al., 2021). The results are so impressive that we show them in Figure 1.

In Figure 1, the colour-coding indicates the result of clustering. The red plots sparsely shown with the titles of the papers are the top-cited papers. Comparing the visualization results of a language model (Sentence-BERT (Reimers and Gurevych, 2019)) and a network embedding (SEAL (Bowman et al., 2015)), we can observe that the top-cited papers are more concentrated in a network embedding model. The entropy of the top-cited papers is 2.900 for the Sentence-BERT model, while it is 1.742 for SEAL. In other words, the top-cited papers have a bias at the SEAL model more than the Sentence-BERT model by 1.1 points in terms of the number of papers with the highest citations.

Thus, several studies have reported that, in some cases, citation information is more effective than linguistic information in predicting the impact of academic literature. In other words, the model for predicting the impact of academic literature requires the active use of citation information.

#### 3.2 Fusion Transformer Model of Linguistic and Citation Information

This study constructs a model that can learn end2end by fusing linguistic and citation information from various academic literature data. Therefore, as shown in Figure 2, we propose the method. The method uses a multilayer perceptron layer (MLP) to fuse the network and the Transformer model for language processing to learn future top-cited papers classification problems. We use Graph-BERT (Zhang et al., 2020) as the Transformer for citation network information and Sci-BERT (Beltagy et al., 2019) as the Transformer for linguistic information. In the previous section 3.1, we found a significant bias between the

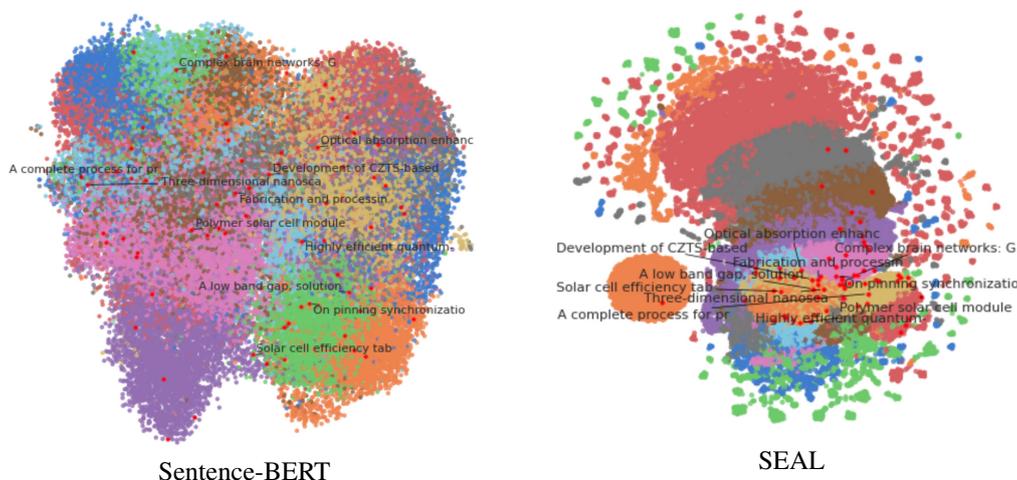


Figure 1: Visualization results of the acquired distributed representation(Ochi et al., 2021). Color coding is the result of the K-means method.

BERT model based on linguistic information and the Embedding model based on networks about the distribution of the papers with the highest number of future citations. Therefore, we considered that only one of the two types of information might be helpful for classification, so we used a parallel model for both rather than a multilayered model in which the output of SciBERT is input to Graph-BERT. We expect to affect, depending on the classification problem, the information via SciBERT is more critical when the linguistic information is valid, and the information via Graph-BERT is more important when the network information is valid. By fusing citation information, we can apply our model even when not all nodes in the network have language information.

In Figure 2, we first select a target paper. The target papers are randomly sampled nodes from the citation network. The proposed method learns and predicts a classification task to determine whether the target papers will likely be the top-cited papers in the future. First, we input three features into the Graph-BERT part. Personalized PageRank (PPR)(Page et al., 1999), Weisfeiler-Lehman Embedding(Niepert et al., 2016), and Hop Distance. Personalized PageRank is a personalized PageRank that computes the PageRank score customized for the target node for all nodes in the network. We order the nodes in decreasing order of PPR value, like a sequence of tokenized words in BERT. We compute Weisfeiller-Lehman Embeddings and input them as features for the aligned nodes. The Hop Distance is the shortest path length in the network from the target node and is input as a feature of the aligned nodes. Next, in Figure 2, we input two pieces of information to Pre-trained SciBERT: the title and abstract of the target paper. We tokenized each and input them as a

series of words, as in BERT.

We only use the [CLS] token, the classification token prefixed at the input of BERT, in the output of Graph-BERT and Pretrained SciBERT. This token allows for efficient training of the classification task. Finally, through the MLP layer, we combine the three [CLS] tokens to learn and predict the classification task of whether the target papers are probably the top-cited papers in the future.

## 4 EXPERIMENT

This section describes the experiments conducted to evaluate our proposed method. First, we describe the seven small datasets of scientific and technical literature we have prepared for our experiments. Next, we train and evaluate our proposed model using a citation classification task. For this purpose, we describe the methods we compare and detail the learning and evaluation conditions.

### 4.1 Scientific Literature Dataset

The data used are seven small datasets extracted from the Web of Science<sup>1</sup> with specific queries. All datasets were for articles published up to 2013. We present an overview of each dataset we extracted in Table 1. The dataset name indicates the representative query in each dataset. In the table, “Num. Articles” indicates the number of articles retrieved in Web of Science by the query. Each dataset uses abstract information as linguistic information and citation information as network information. In the table, from

<sup>1</sup>Web of Science <https://www.webofknowledge.com>

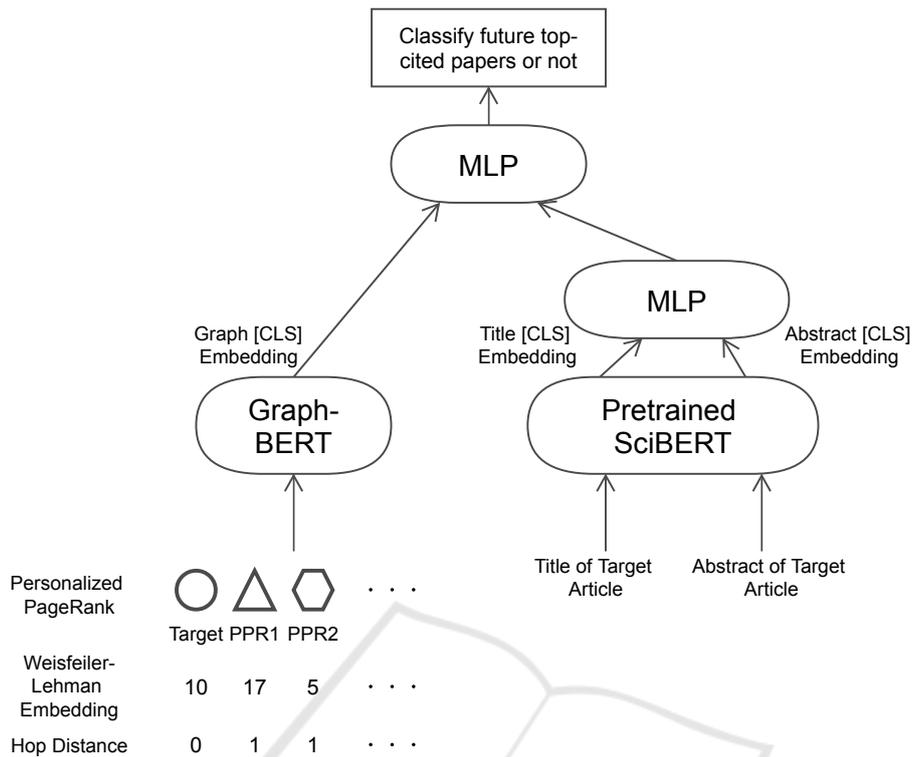


Figure 2: Proposed Method.

“Num. Nodes” to “Gini coef. of Degree Dist.” represent the characteristics of network information and from “Num. Abstracts” to “Word Perplexity” represent the characteristics of linguistic information. In particular, the “Num. Nodes” indicates the number of papers in the network, including papers appearing in the citation information. In contrast, the “Num. Abstracts” is small, indicating that abstract information does not exist in all nodes (papers) in the network.

## 4.2 Classification Problem Setup and Conditions

We consider positive cases as those papers published in 2013 from the dataset extracted in the previous section that is in the top 20% of citations after five years and negative cases as those that are not. We also randomly selected 70% of the papers in our dataset for training and the remainder for evaluation. The training was 50 Epochs, and we calculated Precision, Recall, and F-value as the classification results of the evaluated data in the trained model. We selected three methods for comparison: Graph-BERT only, SciBERT only, and the proposed method. We chose only Graph-BERT, SciBERT, and the proposed method for comparison because Graph-BERT and SciBERT are elements of the proposed method, and the proposed

method is a combined model of the two. We also used one MLP layer and softmax for the classification output. We used publicly available pretrained SciBERT models<sup>2</sup> and performed fine-tuning on each dataset.

## 5 RESULT AND DISCUSSION

We show the Precision, Recall, and F-value results of the classification for each dataset in the table 2. “Graph-BERT”, “SciBERT” and “Proposed Method” in the table represent each method. The “**Bold Characters**” in the result values represents the method with the best result for each dataset and evaluation index. The bottom “Average” represents the simple average of the results per dataset and evaluation index for each method.

First, the “Average” result, which is the average of all F-values, is 0.8349 for the proposed method, 0.8084 for Graph-BERT, and 0.7741 for SciBERT. In other words, the proposed method improves the classification results by 2.65 points over Graph-BERT and 6.08 points over SciBERT. However, when we check the results for each dataset, we find that the proposed method only performs best on the “blackhole” and

<sup>2</sup>SciBERT: <https://github.com/allenai/scibert>

Table 1: Network and linguistic features for each dataset.

Dataset	blackhole	distributed-source-coding	fixed-point	rats	rock-mechanics	taphonomy	thermoelasticity
Num. Articles	1,140	437	3,144	21,226	5,336	3,565	2,184
Num. Nodes	25,211	5,584	32,137	358,035	96,416	140,363	29,776
Num. Edges	50,084	7,586	74,688	613,214	126,961	184,053	45,203
Network Density (%)	0.01576	0.048666	0.014464	0.000957	0.002732	0.001868	0.010197
Avg. Degree	3.973186	2.717049	4.6481	3.425442	2.633609	2.622529	3.036204
Gini Coeff. of Degree Dist.	0.698579	0.592869	0.720047	0.655135	0.586103	0.601066	0.628169
Num. Abstracts	1,097	415	3,009	11,448	4,427	2,617	1,547
Word Perplexity	313.375	110.785	176.444	2341.381	1450.808	1489.115	273.999

“rock-mechanics” datasets. SciBERT shows the best results on the other datasets. In contrast, the GraphBERT model performs best on the other datasets, and SciBERT performs best on the “blackhole” dataset. However, none of the data showed inferior results for the proposed method, and as per the model of the proposed method, we observed a tendency for GraphBERT to give good results and for SciBERT to give close to good results.

What characteristics of the datasets influence the difference in the models that show promising results for each dataset? To clarify this point, we calculated the correlation coefficients between the classification results of the F-values of each dataset (Table 2) and the features of the dataset (Table 1). We show the results in Table 3. The “Feature” column in the table indicates features. Values in “**Bold Characters**” in the table indicate relatively strong correlations with absolute correlation coefficients of 0.6 or more. According to the results, there were no items with a significant correlation between GraphBERT and the feature set used in this study. However, SciBERT showed that the lower the “network density”, the better the classification results. This result means that GraphBERT is impractical when the network density is extremely low, and SciBERT improves the results by predicting only the abstract information. Additionally, the larger the “Avg. Degree” and “Gini Coeff. of Degree Dist.”, the more significant the correlation with SciBERT. The classification problem used in this study was to classify whether the papers published in 2013 would have the highest number of citations by 2018, using papers published up to 2013. In other words, the papers with the highest order were not papers published in 2013 but papers published before that date. Hence, the information that GraphBERT collects from papers published in 2013 is not the papers cited by the cited papers but those papers cited by the corresponding papers. Therefore, GraphBERT does not work well, and SciBERT tends to give better results.

## 6 CONCLUSION

In this paper, we propose a model that fuses linguistic and citation information of academic literature using the Transformer model. The proposed model was trained and evaluated on seven datasets extracted from the Web of Science and showed an average improvement in F-values of 2.65 points over the GraphBERT model alone and 6.08 points over the Scibert model alone. However, some results for individual datasets showed that the single model performed better, indicating that, in many cases, the proposed method tends to produce results comparable to those of the single model that performed better. Correlation analysis of the relationship between the dataset and each model’s F-value and dataset features shows a significant negative correlation between network density and SciBERT results. This result indicates that GraphBERT does not work well when the network information is very sparse and that prediction by linguistic information works well.

In any case, our proposed model improves the classification accuracy of the papers with the highest number of citations after five years. Therefore, the proposed model automatically selects when citation network information is valid and when linguistic information is valid. We conclude that we developed an end2end model that fuses linguistic features and a citation network of scholarly literature data.

However, our proposed method has some limitations. We could not sufficiently clarify whether network or linguistic information is more effective for future top-cited papers classification with the correlation analysis. Therefore, we cannot say that the interaction between linguistic information and the citation network is sufficiently compelling. Additionally, the dataset applied in this study is relatively small. It is necessary to verify whether this method is effective for larger datasets. In the future, we would like to increase the number of features, analyze the conditions under which the model works effectively and present

Table 2: Classification Results.

Dataset	Graph-BERT			SciBERT			Proposed Method		
	Precision	Recall	F-value	Precision	Recall	F-value	Precision	Recall	F-value
blackhole	0.8065	0.5435	0.6494	0.8333	0.9783	0.9000	<b>0.8364</b>	<b>1.0000</b>	<b>0.9109</b>
distributed-source-coding	0.7143	1.0000	0.8333	0.6667	0.4000	0.5000	<b>0.7143</b>	<b>1.0000</b>	<b>0.8333</b>
fixed-point	0.9383	<b>0.9870</b>	0.9620	<b>0.9615</b>	0.974	<b>0.9677</b>	0.9583	0.8961	0.9262
rats	0.8745	<b>0.9806</b>	0.9245	<b>0.9112</b>	0.7737	0.8368	0.8792	<b>0.9562</b>	0.9161
rock-mechanics	<b>0.6512</b>	0.7568	0.7000	0.6327	0.8378	0.7209	0.6271	<b>1.0000</b>	<b>0.7708</b>
taphonomy	0.6727	0.8043	0.7327	<b>0.7115</b>	0.8222	<b>0.7629</b>	0.6727	<b>0.8222</b>	0.7400
thermoelasticity	0.7660	<b>0.973</b>	<b>0.8571</b>	<b>0.8846</b>	0.6216	0.7302	0.7368	0.7568	0.7467
Average	0.7748	0.8636	0.8084	<b>0.8002</b>	0.7725	0.7741	0.7750	<b>0.9188</b>	<b>0.8349</b>

Table 3: The results of comparing the correlation coefficients between the F value of the classification result and each feature for each dataset.

Feature	Method		
	Graph-BERT	SciBERT	Proposed Method
$\log(\text{Num.Nodes})$	0.0421	0.4282	-0.0344
$\log(\text{Num.Edges})$	0.1049	0.5595	0.1012
$\text{NetworkDensity}(\%)$	0.0933	<b>-0.6028</b>	0.1528
Avg. Degree	0.3860	<b>0.8082</b>	<b>0.8158</b>
$\text{GiniCoeff.ofDegreeDist.}$	0.3153	<b>0.8434</b>	<b>0.8029</b>
$\log(\text{Num.Abstacts})$	0.2715	0.5307	0.1552
$\text{Num.Abstacts/Num.Nodes}$	0.5127	0.0138	0.3992
$\text{WordPerplexity}$	0.0080	0.1060	-0.0647

the results more objectively by increasing datasets. Additionally, we would also like to evaluate the integration of methods such as ViT(Dosovitskiy et al., 2021) since there is information on figures and tables in the academic literature data. Also, since our model is a combination of the Transformer model, scalability is expected. We wanted to test the effectiveness of the proposed method on a larger dataset. Since the proposed model is an end2end model, we can quickly increase the number of tasks. We want to test the effectiveness of the proposed method not only in the citation count classification but also for multiple tasks.

## ACKNOWLEDGEMENT

This article is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and supported by JSPS KAKENHI Grant Number JP12345678.

## REFERENCES

Acuna, D. E., Allesina, S., and Kording, K. P. (2012). Predicting scientific success. *Nature*, 489(7415):201–202.

- Ayaz, S., Masood, N., and Islam, M. A. (2018). Predicting scientific impact based on h-index. *Scientometrics*, 114(3):993–1010.
- Bai, X., Zhang, F., and Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407 – 418.
- Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: Pre-trained language model for scientific text. In *EMNLP*.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Cao, X., Chen, Y., and Liu, K. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, 10(2):471 – 484.
- Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. In *ACL*.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Garfield, E. and Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3):195–201.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.
- Miró, Ò., Burbano, P., Graham, C. A., Cone, D. C., Ducharme, J., Brown, A. F. T., and Martín-Sánchez, F. J. (2017). Analysis of h-index and other bibliometric markers of productivity and repercussion of a selected sample of worldwide emergency medicine researchers. *Emergency Medicine Journal*, 34(3):175–181.
- Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2014–2023, New York, New York, USA. PMLR.
- Ochi, M., Shiro, M., Mori, J., and Sakata, I. (2021). Which is more helpful in finding scientific papers to be top-cited in the future: Content or citations? case analysis in the field of solar cells 2009. In Mayo, F. J. D., Marchiori, M., and Filipe, J., editors, *Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST 2021, October 26-28, 2021*, pages 360–364. SCITEPRESS.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.
- Park, I. and Yoon, B. (2018). Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics*, 12(4):1199–1222.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Sasaki, H., Hara, T., and Sakata, I. (2016). Identifying emerging research related to solar cells field using a machine learning approach. *Journal of Sustainable Development of Energy, Water and Environment Systems*, 4:418–429.
- Schreiber, M. (2013). How relevant is the predictive power of the h-index? a case study of the time-dependent hirsch index. *Journal of Informetrics*, 7(2):325 – 329.
- Stegehuis, C., Litvak, N., and Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, 9.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Yan, E. and Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, 8(2):295–309.
- Yi, Z., Ximeng, W., Guangquan, Z., and Jie, L. (2018). Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology. *Proceedings of the Association for Information Science and Technology*, 55(1):598–607.
- Zhang, J., Zhang, H., Xia, C., and Sun, L. (2020). Graphbert: Only attention is needed for learning graph representations. *CoRR*, abs/2001.05140.