BERT-Based Scientific Paper Quality Prediction

Taiki Sasaki[1], Yasuaki Ito[1](✉), Koji Nakano[1], and Akihiko Kasagi[2]

[1] Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima, Japan
{taiki,yasuaki,nakano}@cs.hiroshima-u.ac.jp
[2] Fujitsu Ltd., Tokyo, Japan
kasagi.akihiko@fujitsu.com

Abstract. In recent years, scholarly databases have made many scientific papers available on the Internet. While these databases facilitate access to excellent papers, they also increase the possibility of encountering inferior papers. However, it is difficult to predict the quality of a paper just from a glance at the paper. In this paper, we propose a machine learning approach to predicting the quality of scientific papers. Specifically, we predict the quality of an article by classifying for the abstract of the paper whether the article is included in a superior journal or not. The proposed model is trained using a BERT-based model widely used in natural language processing. After training, we achieved a test accuracy of 95.1% and 89.6% in medicine and computer science, respectively. In addition, the results of the classification are visualized by evaluating the sentence combinations in the abstract to clarify the details of the classification.

Keywords: Paper quality prediction · Machine learning · BERT

1 Introduction

Text classification is a fundamental problem in natural language processing (NLP), and it has been applied in various fields such as translation, dialogue response, sentiment analysis, and summarization. In recent years, machine learning models have been widely used for text classification [20]. In these approaches, the text is input to a machine learning model and it is trained to classify the text. In NLP using machine learning, recurrent neural networks (RNNs) such as long short-term memory (LSTM) with recursive structures have often been used in natural processing using machine learning. However, these models require sequential processing from the beginning to the end of a sentence, which prevents parallel computation. This is a critical drawback for training networks, which generally require a large amount of time. For this problem, the Transformer was proposed [22]. By introducing the self-attention structure, Transformer can achieve the same or better performance as RNNs without recursive structure.

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13532, pp. 212–223, 2022.
https://doi.org/10.1007/978-3-031-15937-4_18

The Transformer processes the inputs simultaneously and computes the attention weights among them. This allows the network to be trained on large data sets using parallel processing. Also, BERT (Bidirectional Encoder Representations from Transformers) [13] utilizing Transformer technology is one of the most successful models currently available; the performance of natural language processing using machine learning has been greatly improved by BERT.

The main contribution of this paper is to propose a method for predicting the quality of scholarly papers using machine learning. Recently, many scientific papers have been available on the Internet through PubMed [4], Web of Science [7], Google Scholar [2], and others. While they have made it easier to browse superior papers, they have also increased the chances of encountering inferior papers. Generally, the number of citations is used as an indicator of the superiority of a paper. However, it is not easy to predict quality simply by the number of citations alone, as it is highly dependent on the time of publication, and moreover, it is impossible to predict a paper before submission. Therefore, in this paper, we consider papers published in superior journals to be superior papers and papers published in less superior journals to be less superior papers. Furthermore, we consider the abstracts of papers published in these journals to be similar, thus predicting the quality of papers based on the abstracts. Based on the above idea, in the proposed approach, the quality prediction of papers is obtained as a classification problem for the abstracts of papers. Specifically, the proposed method uses a BERT-based model to classify whether an article is included in the upper or the lower-ranked journal in the Average Journal Impact Factor Percentile [1] from its abstract. In this paper, we show that different datasets of pre-training of the proposed BERT-based model affect classification accuracy. The results of training the models show that the models can classify whether the input abstracts are from superior or less superior journals with a test accuracy of 95.1% and 89.6% in the field of medicine and computer science, respectively.

As related work, studies predicting the quality of academic papers have been conducted [8,16]. In those studies, there are mainly two types of approaches of prediction. One is to predict the quality of a paper based only on its content, using the title, abstract, text, figures, tables, references, and appearance of the paper [10,14,19,21,24]. The other is to estimate the quality based on the contents of the paper as well as additional information outside of the paper, such as author reputations, impact factor of the journal in which the paper was published, cited papers and the citation network composed of them [9,11,12,17,25]. Unlike the above methods, the proposed method classifies whether or not the content of an abstract is that of a superior journal. This classification is inspired by the idea that papers in a good journal have well-described abstracts. The proposed approach does not aim to judge the excellence of the research, but focuses on the quality of writing of the papers.

The rest of this paper is organized as follows. We briefly introduce BERT in Sect. 2. In Sect. 3, we show our proposed machine learning approach using the BERT-based model, and experimental results are presented in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 BERT

Bidirectional Encoder Representations from Transformers (BERT) [13] is one of the state-of-the-art machine learning techniques based on Transformer [22] for NLP. The technique includes model structures and learning approaches. In this section, we briefly introduce the technique.

Figure 1 illustrates an outline of the structure of the BERT model. Given an input sequence of N words, each word is converted to a token and then each of them is mapped to a vector data of size k by word embedding. After that, L transformer encoder blocks transform the tokens to contain more accurate information. Each Transformer encoder block based on the encoder of the Transformer has A attention heads and H hidden layers. The BERT model has configurations of various sizes, and typical models and their number of parameters are shown in Table 1. Users can choose whether to use all of the output of the last Transformer encoder block or the part of one for final classification. Usually, it is sufficient to use only the first output for classification tasks, and the final classification result is obtained using the classifier based on the first output as shown in the figure.

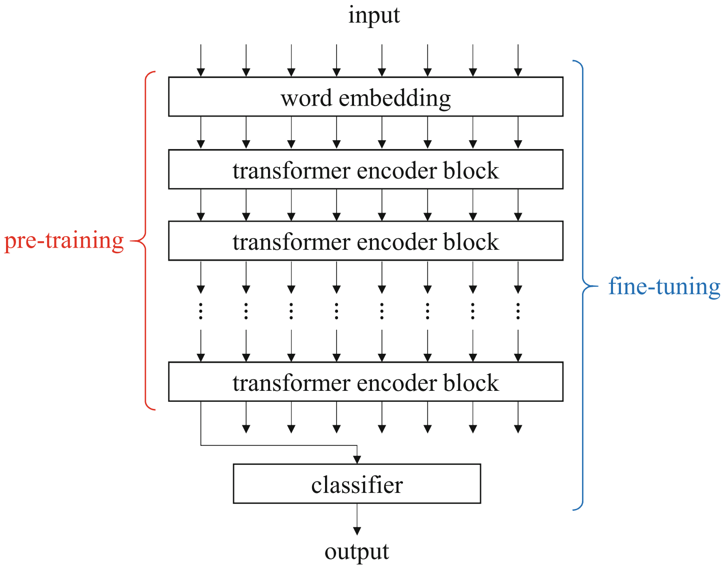
input
word embedding
transformer encoder block
transformer encoder block
pre-training
transformer encoder block
fine-tuning
classifier
output

Fig. 1. Structure of the BERT model

The training of BERT models consists of two phases: *pre-training phase* and *fine-tuning phase*. In general, the pre-training phase involves training the model on a large corpus such as Wikipedia. On the other hand, in the fine-tuning phase, the weight parameters obtained in pre-training are used as initial values for the model, and training is performed on the target task. This phase can often be completed with much less computation than the pre-training phase. We give an explanation about these phases as follows.

Table 1. Model configuration of BERT models

Model	L	H	A	Parameters
BERT-tiny	2	128	2	4,386,050
BERT-mini	4	256	4	11,170,817
BERT-small	4	512	8	28,764,161
BERT-medium	8	512	8	41,373,697
BERT-base	12	768	12	109,483,009

The network training in the pre-training phase is unsupervised and consists of two tasks, *masked language model* (MLM) and *next sentence prediction* (NSP). In this training, we train only on word embedding and transformer encoder blocks. In MLM, we input sentences of tokens with some of them masked to train the model to predict the original tokens. In NSP, we input two sentences concatenated to train the model to predict whether the two sentences are consecutive or not. When selecting two sentences, 50% of the input is actually consecutive sentences and the other 50% is randomly selected from the data set. By training on a large number of sentences for these two tasks, we obtain a model that can capture the features of the sentences. Since the computational cost of pre-training is enormous, we can utilize BERT models already trained on massive corpora such as Wikipedia, BookCorpus, and MEDLINE/PubMed [6,23], and often employ these models for the following fine-tuning phase.

In the fine-tuning phase, we train the whole network on the target task. The model obtained in the above pre-training is used as the initial values and trained as supervised learning. In general, this training requires fewer iterations than the pre-training. In the proposed method, the pre-trained model is trained as a classification problem.

3 Proposed Quality Prediction of Scientific Papers

This section presents the proposed method, the BERT model for predicting the quality of papers as a classification problem, the utilized dataset, and the classifier.

3.1 Dataset of Scientific Papers

In this work, we use the semantic scholar open research corpus (S2ORC) [5,18] version 20200705v1 as a dataset of scientific papers. S2ORC is used for natural language processing and text mining research. The dataset contains 136M paper data, of which 12M are full-text papers, covering various fields of research. In this study, we employ abstracts of papers in the fields of medicine and computer science from the data set.

3.2 Quality Classification of Papers

In order to predict the quality of papers, we introduce the Average Journal Impact Factor Percentile (Average JIF Percentile) provided by Journal Citation

Reports [1] as the metric of article quality. JIF Percentile is a metric that indicates the top percentage of journals in a given field in terms of impact factor in that field. JIF Percentile is obtained by the following formula [3]:

$$\frac{N - R + 0.5}{N}, \tag{1}$$

where N is the number of journals in the category and R is the descending rank. Average JIF Percentile is the average of the JIF Percentile values of the target fields, which takes into account the fact that the target fields cover multiple fields. In this study, we consider predicting the quality of papers as a classification problem. Let J_U be the set of journals whose Average JIF Percentile is 0.8 or higher, and J_L denote the set of journals whose Average JIF Percentile is 0.2 or lower. In the classification problem, given an abstract of a paper, we classify whether the paper is included in J_U or J_L. The reader might think that if the journals in a particular field are biased toward either of J_U and J_L, then this classification problem would lead to a different classification problem of whether a paper is in a particular field or not. Tables 2 and 3 are the 10 journals with the most papers in J_U and J_L for medicine and computer science, respectively. According to the tables, there is no significant unbalance in the fields of papers included in J_U and J_L, respectively. This means that this classification problem cannot be correctly classified only by finding papers in a specific field.

Table 2. Journals with the most papers for medicine

(a) 10 journals with the most papers in J_U

Journal	#papers	%
Nature Communications	19744	6.67
Nanoscale	17215	5.82
Nutrients	9582	3.24
JAMA Internal Medicine	9535	3.22
JAMA Pediatrics	7813	2.64
Small	7377	2.49
eLife	7188	2.43
PLoS Genetics	6088	2.06
PLoS Neglected Tropical Diseases	6028	2.04
PLoS Pathogens	5871	1.98

(b) 10 journals with the most papers in J_L

Journal	#papers	%
European Journal of Hospital Pharmacy-Science and Practice	3687	3.64
Vascular and Endovascular Surgery	3460	3.41
Natural Product Communications	3162	3.12
Therapeutic Innovation & Regulatory Science	2758	2.72
Indian Journal of Psychiatry	2101	2.07
Psychiatria Danubina	1842	1.82
Journal of Traditional Chinese Medicine	1800	1.77
Indian Journal of Surgery	1699	1.68
Australasian Psychiatry	1660	1.64
Oncology Research and Treatment	1616	1.59

Table 3. Journals with the most papers for computer science

(a) 10 journals with the most papers in J_U

Journal	#papers	%
PLoS Computational Biology	4279	9.23
IEEE Transactions on Industrial Informatics	3113	6.71
IEEE Transactions on Smart Grid	3038	6.55
IEEE Transactions on Information Forensics and Security	2183	4.71
IEEE Transactions on Neural Networks and Learning Systems	2121	4.57
IEEE Transactions on Cybernetics	1994	4.30
IEEE Internet of Things Journal	1990	4.29
IEEE Robotics and Automation Letters	1915	4.13
IEEE Transactions on Robotics	1762	3.80
IEEE Wireless Communications Letters	1730	3.73

(b) 10 journals with the most papers in J_L

Journal	#papers	%
International Journal of Distributed Sensor Networks	3461	15.92
Security and Communication Networks	1997	9.19
Journal of Internet Technology	1612	7.42
Turkish Journal of Electrical Engineering and Computer Sciences	1393	6.41
International Journal on Artificial Intelligence Tools	1007	4.63
Journal of Semiconductor Technology and Science	797	3.67
Electronics and Communications in Japan	754	3.47
Journal of Medical and Biological Engineering	724	3.33
International Journal on Software Tools for Technology Transfer	720	3.31
Cognitive Processing	628	2.89

3.3 BERT-Based Model of Quality Prediction of Scientific Papers

The proposed BERT-based model of quality prediction of scientific papers is the structure shown in Fig. 1. The classifier has an input that corresponds to the first token in the output of the last transformer encoder block and outputs the classification result. The model including the classifier is trained in the fine-tuning phase to output 1 if the input abstract is included in J_U and 0 if it is included in J_L. The classifier uses a fully connected layer with one output channel and a sigmoid function as the activation function. We note that since the problem targeted by the proposed model is to classify peer-reviewed papers, this is a more difficult problem than the classification problem that predicts acceptance and non-acceptance for publishing [10,24].

4 Experimental Results

In this section, we show the methodology for training models for predicting paper quality as the target task, and evaluate the resulting models. In this study, we train models to predict the quality from abstracts for two research fields, medicine and computer science. In the proposed approach, three types of models have been employed as the BERT-models trained in the pre-training phase. Two of them are pre-trained models from the TensorFlow Hub [6], one trained on the Wikipedia and BookCoups datasets, and the other trained on the MEDLINE/PubMed datasets. Please refer [6] for the already pre-trained models on the Wikipedia and BookCoups datasets and the MEDLINE/PubMed datasets. Since the pre-trained model sizes available in the TensorFlow Hub vary by dataset, we experiment with the BERT-tiny, BERT-mini, BERT-small, and BERT-base models for the Wikipedia and BookCoups datasets, and the BERT-base model for the MEDLINE/PubMed datasets. The remaining one is a model that we trained by ourselves using the abstracts of papers in S2ORC [5]. On the other hand, in the fine-tuning phase, we fine-tune the models on abstracts of papers in S2ORC. In the following, we show the details of the training in the pre-training phase and the fine-tuning phase.

4.1 Training in the Pre-training Phase on Abstracts from S2ORC

Here, we show the training of the BERT models on all abstracts in the S2ORC dataset trained from scratch. The BERT-based models were trained using MLM and NSP tasks shown in Sect. 2. Each model was trained for 3,000,000 steps, with a batch size of 8, and a maximum input size 512. The training is optimized by Adam with learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L_2 weight decay of 0.01, learning rate warmup over the first 10,000 steps, and linear decay of the learning rate. We use GELU [15] as the activation function and the sum of the likelihood of MLM and NSP as the training loss.

4.2 Training in the Fine-Tuning Phase

The training in the fine-tuning phase was performed using each model trained in the pre-training phase as initial values of weights. In the experiment, models were trained on S2ORC dataset from two fields, medicine and computer science, respectively. The number of papers in each field of training data and test data is shown in Table 4. Each model was trained for 50 epochs, with a batch size of 64, and a maximum input size 512. The training is optimized by AdamW with a learning rate of 0.00003, L_2 weight decay of 0.01, learning rate warm-up over the first 10,000 steps, and linear decay of the learning rate. We use GELU as the activation function and the binary cross-entropy loss as the training loss.

Table 4. The number of abstracts from S2ORC dataset in the fine-tuning phase

	Medicine		Computer science	
	J_U	J_L	J_U	J_L
Training data	294,840	100,416	47,381	21,737
Test data	1,000	1,000	1,000	1,000

4.3 The Test Accuracy of Prediction of the Trained Model

Table 5 shows the test accuracy of prediction of the trained models for each research field. The table shows that for both fields, the larger models are more accurate, and the model pre-trained on MEDLINE/PubMed dataset is the most accurate. The reason for the lower accuracy when using the S2ORC dataset for pre-training despite having the same scholarly articles as the MEDLINE/PubMed dataset is due to the smaller size of the dataset compared to the other two datasets. As a result of training the model, we achieved a test accuracy of 95.1% in the medical field and 89.6% in the computer science field. As shown in Tables 2 and 3, since there is little unevenness in fields between J_U and J_L journals, this result implies that the model does not classify papers by finding specific research fields from their abstracts, but can perform the classification in terms of abstract presentation.

Table 5. The test accuracy of prediction

Pre-training dataset	Model	Medicine	Computer science
Wikipedia and BookCorpus	BERT-tiny	0.9014	0.8280
	BERT-mini	0.9254	0.8380
	BERT-small	0.9265	0.8519
	BERT-medium	0.9290	0.8610
	BERT-base	0.9304	0.8769
MEDLINE/PubMed	BERT-base	**0.9509**	**0.8955**
S2ORC	BERT-tiny	0.8805	0.8285
	BERT-mini	0.9114	0.8365
	BERT-small	0.9170	0.8455
	BERT-medium	0.9260	0.8530
	BERT-base	0.9370	0.8625

4.4 Detailed Analysis of the Prediction

To clarify the classification details of the proposed model, we performed the classification on subsets of the sentences in the abstract. Specifically, Tables 6 and 7 show the results of the classification performed on the i-th to j-th sentences for two abstracts sampled from J_U and J_L, respectively. We note that these two abstracts are simply sampled from each dataset and are not meant to judge their quality of them. The classification results are the top right values of the table, 0.9736 and 0.0429, respectively, indicating that the model correctly classifies each. Also, the values of the diagonal elements indicate the output values of each sentence evaluated on itself. From the table, it appears that the proposed model outputs the final result by considering multiple sentences, although both abstracts contain sentences with large and small values. Focusing on the output values of single sentences, there are sentences with high or low values, but the final estimated results seem to be evaluated by considering multiple sentences. In addition, the proposed model may be used as a supporting tool when writing papers, since the output as shown in Tables 6 and 7 provides at-a-glance information on good and bad descriptions in abstracts.

Table 6. Detailed analysis for an abstract in J_U sampled from computer science papers in S2ORC [18]

(a) abstract in J_U

1. Deeper neural networks are more difficult to train.
2. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously.
3. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions.
4. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth.
5. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers-8x deeper than VGG nets but still having lower complexity.
6. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set.
7. This result won the 1st place on the ILSVRC 2015 classification task.
8. We also present analysis on CIFAR-10 with 100 and 1000 layers.
9. The depth of representations is of central importance for many visual recognition tasks.
10. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset.
11. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

(b) Output values of the model for the i-th to j-th sentences

							j					
		1	2	3	4	5	6	7	8	9	10	11
	1	0.9494	0.9146	0.9662	0.9832	0.9221	0.9062	0.8776	0.9218	0.9732	0.9837	0.9736
	2		0.9128	0.9336	0.9424	0.9250	0.9306	0.9035	0.9492	0.9766	0.9873	0.9740
	3			0.9324	0.9837	0.9020	0.8694	0.8283	0.8847	0.9532	0.9764	0.9623
	4				0.9762	0.6459	0.5470	0.4694	0.6045	0.7709	0.8799	0.8965
	5					0.3641	0.2645	0.1109	0.1645	0.2636	0.6235	0.8059
i	6						0.3372	0.1311	0.2202	0.3566	0.6363	0.7867
	7							0.3083	0.3480	0.3527	0.8034	0.8298
	8								0.1070	0.4168	0.8924	0.7935
	9									0.9597	0.9852	0.8694
	10										0.9813	0.8248
	11											0.9170

Table 7. Detailed analysis for an abstract in J_L sampled from computer science papers in S2ORC [18]

(a) abstract in J_L

1. GPUs are one of the most prevalent platforms for accelerating general-purpose workloads due to their intuitive programming model, computing capacity, and cost-effectiveness.
2. GPUs rely on massive multi-threading and fast context switching to overlap computations with memory operations.
3. Among the diverse GPU workloads, there exists a class of kernels that fail to maintain a sufficient number of active warps to hide the latency of memory operations, and thus suffer from frequent stalling.
4. We observe that these kernels will benefit from increased levels of Instruction-Level Parallelism and we propose a novel architecture with lightweight Out-Of-Order execution capability.
5. To minimize hardware overheads, we carefully design our extension to highly re-use the existing micro-architectural structures.
6. We show that the proposed architecture outperforms traditional platforms by 15 to 46 percent on average for low occupancy kernels, with an area overhead of 0.74 to 3.94 percent.
7. Finally, we prove the potential of our proposal as a GPU u-arch alternative, by providing a 5 percent speedup over a wide collection of 63 general-purpose kernels with as little as 0.74 percent area overhead.

(b) Output values of the model for the i-th to j-th sentences

				j			
	1	2	3	4	5	6	7
1	0.0209	0.0159	0.0082	0.0054	0.0044	0.0116	0.0429
2		0.0700	0.1192	0.0123	0.0064	0.0110	0.0497
3			0.0188	0.0084	0.0055	0.0115	0.0110
i 4				0.9172	0.4710	0.3974	0.3587
5					0.2333	0.4810	0.4205
6						0.7966	0.4325
7							0.4937

5 Conclusions

In this paper, we proposed a method for predicting the quality of scholarly papers using machine learning. We predict the quality of papers as a classification problem for the abstracts of papers whether the paper is included in superior journals or less superior journals. We used BERT-based models and trained them on several datasets. As a result of our experiment, we showed that different datasets of pre-training of the proposed BERT-based model affect classification accuracy. Also, the results of training the models showed that the models could classify whether the input abstracts are from superior or less superior journals with a test accuracy of 95.1% and 89.6% in the field of medicine and computer science, respectively. Furthermore, by evaluating the sentence combinations in the abstracts, we clarified the details of the classification results and visualized them.

References

1. Journal Impact Factor: Journal Citation Reports Science Edition (Clarivate Analytics 2021). https://jcr.clarivate.com/
2. Google Scholar. https://scholar.google.co.jp/
3. Journal Impact Factor Percentile. https://help.incites.clarivate.com/incitesLiveJCR/glossaryAZgroup/g8/9586-TRS.html
4. PubMed. https://pubmed.ncbi.nlm.nih.gov/
5. S2ORC. https://github.com/allenai/s2orc
6. TensorFlow Hub. https://tfhub.dev/
7. Web of Science. https://clarivate.jp/training/web-of-science/
8. Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., Xia, F.: An overview on evaluating and predicting scholarly article impact. Information **8**(3), 73 (2017). https://doi.org/10.3390/info8030073
9. Bai, X., Zhang, F., Lee, I.: Predicting the citations of scholarly paper. J. Inform. **13**(1), 407–418 (2019). https://doi.org/10.1016/j.joi.2019.01.010
10. Maillette de Buy Wenniger, G., van Dongen, T., Aedmaa, E., Kruitbosch, H.T., Valentijn, E.A., Schomaker, L.: Structure-tags improve text classification for scholarly document quality prediction. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 158–167 (2020). https://doi.org/10.18653/v1/2020.sdp-1.18, https://aclanthology.org/2020.sdp-1.18
11. Castillo, C., Donato, D., Gionis, A.: Estimating number of citations using author reputation. In: Proceeding of the 14th International Conference on String Processing and Information Retrieval, pp. 107–117 (2007)
12. Davletov, F., Aydin, A.S., Cakmak, A.: High impact academic paper prediction using temporal and topological features. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 491–498 (2014)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
14. van Dongen, T., Maillette de Buy Wenniger, G., Schomaker, L.: SChuBERT: scholarly document chunks with BERT-encoding boost citation count prediction. In: Proceedings of the First Workshop on Scholarly Document Processing. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.sdp-1.17
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs) (2020)
16. Hou, J., Pan, H., Guo, T., Lee, I., Kong, X., Xia, F.: Prediction methods and applications in the science of science: a survey. Comput. Sci. Rev. **34**, 100197 (2019). https://doi.org/10.1016/j.cosrev.2019.100197
17. Livne, A., Adar, E., Teevan, J., Dumais, S.: Predicting citation counts using text and graph mining. In: iConference 2013, Workshop on Computational Scientometrics: Theory and Application, February 2013
18. Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.: S2ORC: the semantic scholar open research corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4969–4983. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.447
19. Ma, A., Liu, Y., Xu, X., Dong, T.: A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics **126**(8), 6803–6823 (2021). https://doi.org/10.1007/s11192-021-04033-7

20. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Netw. Learn. Syst. **32**(2), 604–624 (2021). https://doi.org/10.1109/TNNLS.2020.2979670
21. Shen, A., Salehi, B., Qi, J., Baldwin, T.: A multimodal approach to assessing document quality. J. Artif. Intell. Res. **68**, 607–632 (2020). https://doi.org/10.1613/jair.1.11647
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
24. Yang, P., Sun, X., Li, W., Ma, S.: Automatic academic paper rating based on modularized hierarchical convolutional neural network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2: Short Papers, pp. 496–502 (2018). https://doi.org/10.18653/v1/P18-2079
25. Zhao, Q., Feng, X.: Utilizing citation network structure to predict paper citation counts: a deep learning approach. J. Inform. **16**(1), 101235 (2022). https://doi.org/10.1016/j.joi.2021.101235