

Chinese paper classification based on pre-trained language model and hybrid deep learning method

Xin Luo, Sofianita Mutalib, Syaripah Ruzaini Syed Aris

School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received May 6, 2024

Revised Sep 11, 2024

Accepted Sep 30, 2024

Keywords:

Bidirectional encoder representations from transformers

Chinese scientific literature dataset

Deep learning model

Model combination

Paper classification

Pre-training language model

ABSTRACT

With the explosive growth in the number of published papers, researchers must filter papers by category to improve retrieval efficiency. The features of data can be learned through complex network structures of deep learning models without the need for manual definition and extraction in advance, resulting in better processing performance for large datasets. In our study, the pre-trained language model bidirectional encoder representations from transformers (BERT) and other deep learning models were applied to paper classification. A large-scale chinese scientific literature dataset was used, including abstracts, keywords, titles, disciplines, and categories from 396 k papers. Currently, there is little in-depth research on the role of titles, abstracts, and keywords in classification and how they are used in combination. To address this issue, we evaluated classification results by employing different title, abstract, and keywords concatenation methods to generate model input data, and compared the effects of a single sentence or sentence pair data input methods. We also adopted an ensemble learning approach to integrate the results of models that processed titles, keywords, and abstracts independently to find the best combination. Finally, we studied the combination of different types of models, such as the combination of BERT and convolutional neural networks (CNN), and measured the performance by accuracy, weighted average precision, weighted average recall, and weighted average F1 score.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sofianita Mutalib

School of Computing Sciences, College of Computing, Informatics and Mathematics

Universiti Teknologi MARA

Shah Alam, Selangor 40450, Malaysia

Email: sofianita@uitm.edu.my

1. INTRODUCTION

In recent years, due to the increasing number of scientific papers, researchers need to retrieve papers related to their research fields more efficiently. The category labeling of scientific papers is a task that must be completed in document taxonomy. If it is completed by manpower, professional knowledge must be required, which is costly and inefficient. It is an important task in natural language processing (NLP) to complete the automatic classification of papers through machine learning algorithms and achieve practical accuracy. Applying traditional machine learning algorithms to classify papers requires completing two steps. First, obtain the document representation vector through term frequency-inverse document frequency (TF-IDF), Word2Vec, global vectors for word representation (GloVec), FastText, and other methods, and then use them as input data for classification algorithms such as naive Bayes, decision tree, support vector machine, and neural network [1]–[4].

Traditional machine learning classifiers have high accuracy and efficiency in small data sets, but struggle with large-scale data sets with complex features. Deep learning offers advantages like avoiding manual feature definition and using complex network structures to extract and generalize data features, resulting in higher document classification accuracy. Convolutional neural networks (CNN) can extract local features and transfer them to global features by pooling layers from text sequences [5], [6]. Recurrent neural network (RNN) and long short-term memory (LSTM) models are suitable for processing sequential data because they can remember the dependency between tokens [7], [8]. Bidirectional encoder representations from transformers (BERT) is a bidirectional language model that performs mask language model (MLM) tasks and next sentence prediction (NSP) tasks on large-scale corpora to extract contextual semantic information and obtain relationships between sentences [9]–[11]. BERT can better extract sentence feature information and understand the semantic relationships between sentences, due to the use of bidirectional encoding and self-attention mechanisms.

The title, abstract, and keywords of scientific papers are the most important meta-information of the paper, including semantic information that can be used to distinguish different categories. They can be easily obtained as training corpus. Currently, there is little in-depth research on the role of titles, abstracts, and keywords in classification and how they are used in combination. In our study, we applied the pre-trained language model BERT and other deep learning models such as CNN and LSTM for paper classification. We used different combinations of input data features and models, and measured performance through accuracy, weighted average precision, weighted average recall, and weighted average F1-score. This paper is organized as follows. We first introduce the latest research progress in paper classification, then introduce the research methods including datasets and experimental settings, then discuss the experimental results, and finally conclude.

2. RELATED WORK

Classification of scientific papers using machine learning methods has been extensively studied, and most of the research uses the metadata in the papers, that is, to extract the feature vectors of the papers from the title, abstract, keywords, and other information to train various classifiers. Several researchers [12]–[14] proposed to use support vector machine classifier or Bayesian algorithm to realize the classification of papers. Xiaohua and Haiyun [15] proposed a hierarchical classification method for Chinese scientific papers based on important words in titles, keywords, and abstracts. Words in the paper text will also be used if they have high mutual information value with important words. For words in different paper areas, a β value is assigned to the feature vector calculation formula, and the values are arranged in the following order title > summary > keywords > main text.

Using deep learning techniques to classify scientific papers has become popular in recent years. Chouyyekh *et al.* [16] proposed to use CNN to classify scientific papers, and used the "Web of Science Dataset" as an experimental dataset, which contains input text sequence, target label value, domain, keywords, and summary information of 35238 papers. Burns *et al.* [17] built deep learning models for evidence classification from the open-access biomedical literature, developed a large-scale corpus from PubMed and PubMed central open-access records and then used Glove, FastText, and ELMo algorithms to learn word embedding. They also use CNN, LSTM, and attention mechanisms to improve the effect of classification [17]. Samami and Soure [18] used ensemble deep learning models to classify Lupus scientific articles, namely, a combination of LSTM, cuda deep neural network gated recurrent unit (CuDNNGRU), RNN, and CNN models were used to classify paper abstracts, and the final classification results were selected through voting, and the results showed that the ensemble method improves the reliability of classification [18]. Bogdanchikov *et al.* [19] used a deep learning model and naive Bayes algorithm to classify scientific papers written in Kazakh language, and processed image and text separately. The experimental results showed that the accuracy was improved by using multimodal information compared to using text features or images alone. Semantic featured convolution neural networks (SF-CNN) were proposed in [20] to improve the performance of traditional CNN which does not consider the semantics of bag-of-words. The training dataset was collected from ArXiv, and experimental results showed that the classification accuracy reached 94%.

For research on the classification of Chinese scientific papers, Lili *et al.* [21] used the BERT model to classify different types of Chinese literature and achieved a classification accuracy of 76.95% and 68.55% respectively [21]. Another study also showed that BERT models outperformed the support vector machine model, among which the BERT-re-pretraining-med-Chinese model performed best [22]. Hongling *et al.* [23] studied the impact of stop words in scientific papers on classification performance, and they found that RNN, LSTM, and gated recurrent unit (GRU) models could achieve better performance without removing stop words. Using Adam or stochastic gradient descent (SGD) optimizer for RNN and LSTM models, and Adadelata or SGD optimizer for GRU models can improve the classification effect [23]. Jie [24] developed an automatic

document classification system that uses the skim-gram word embedding model to extract the feature matrix of the document and adopts the CNN model as the classifier. The first-level, second-level, and final-level classification accuracy were 97.66%, 95.12%, and 92.42% respectively [24]. Zhang *et al.* [25] studied the role of the full-text and structural information of papers in classification. The use of the pre-trained model LongFormer showed that the introduction of full-text information will lead to a decrease in classification accuracy, while abstract, keyword, and title information plays a decisive role in paper classification [25].

3. METHOD

We first downloaded the public data Chinese scientific literature (CSL) dataset and preprocessed it to generate training datasets, development datasets, and test datasets. Then we designed four types of experiments, including different combinations of keywords, titles, and abstracts in single-sentence mode, single-sentence and sentence-pair input methods, ensemble learning methods, and different model combinations, to compare their classification performance under different parameter settings. Classification performance is measured by accuracy, precision, recall, and F1 score.

3.1. The dataset

A large-scale CSL dataset was built in [26], which contains the titles, abstracts, keywords, and other fields of 396 k academic papers, it should be the first public CSL dataset. It can be used for NLP tasks such as text summarization, keyword generation, and text classification. The paper's meta-information is from the National Engineering Research Center for Science and Technology Resources Sharing Service (NSTR), dated from 2010 to 2020. The characteristics of the dataset include wider discipline coverage, new data source (because most of the current paper data sets are taken from arXiv, PubMed, ACLAnthology, and MAG), higher quality and accuracy (the paper has been peer-reviewed). The statistical information of the entire data set is shown in Table 1. Since public Chinese scientific paper datasets are rare, most researchers obtain the meta-information of papers from the library literature sharing platform or online literature databases, and datasets constructed by them are generally not made public, so the CSL data set can provide a benchmark for performance evaluation of Chinese scientific paper classification models. A sample from this dataset is shown in Table 2.

Table 1. Detailed statistics of the CSL dataset

Category	#d	Len(T)	Len(A)	num(K)	#samples	Discipline examples
Engineering	27	19.1	210.9	4.4	177,600	Mechanics, Architecture, Electrical science
Science	9	20.7	254.4	4.3	35,766	Mathematics, Physics, Astronomy, Geography
Agriculture	7	17.1	177.1	7.1	39,560	Crop science, Horticulture, Forestry
Medicine	5	20.7	269.5	4.7	36,783	Clinical medicine, Dental medicine, Pharmacy
Management	4	18.7	157.7	6.2	23,630	Business management, Public administration
Jurisprudence	4	18.9	174.4	6.1	21,554	Legal science, Political science, Sociology
Pedagogy	3	17.7	179.4	4.3	16,720	Pedagogy, Psychology, Physical education
Economics	2	19.5	177.2	4.5	11,558	Theoretical economics, Applied economics
Literature	2	18.8	158.2	8.3	10,501	Chinese literature, Journalism
Art	1	17.8	170.8	5.4	5,201	Art
History	1	17.6	181.0	6.0	6,270	History
Strategies	1	17.5	169.3	4.0	3,555	Military science
Philosophy	1	18.0	176.5	8.0	7511	Philosophy
All	67				396,209	

#d: the number of disciplines in the category. len(T): average length of each title; len(A): average length of each abstract; num(K): average number of keywords

Table 2. A sample from the CSL dataset in English

Title	Abstract	Keywords	Discipline	Category
Exploration on improving peasants' scientific and cultural quality by using distance education	Starting from the importance of improving farmers' scientific and technological quality, this paper discusses the ways and characteristics of distance education	Pedagogy; Distance education; Comments; Quality	Agricultural Engineering	Engineering

3.2. Feature extraction

BERT is an open-source machine learning framework for NLP. BERT is designed to help computers understand the meaning of ambiguous language in the text by using surrounding text to establish context. In many natural language understanding (NLU) tasks, such as sentiment analysis, semantic role annotation, and text classification, BERT can achieve better performance than other deep learning methods. The process of

using BERT to generate the token feature representation vector of the text sequence is as follow. For English text sequences, word segmentation is first performed, while for Chinese text sequences, it is processed character by character, and finally forms a sequence composed of English words or Chinese character tokens. [CLS] and [SEP] tags are used to represent the classification and separator of sentences and will be added to the beginning and end of the sentence respectively. The embedding representation of each token is obtained from vocab, a vocabulary provided by BERT. The sum of token embedding, segment embedding, and position embedding is input to the transformer layer, and the token vector with global semantic information and class vector will be generated. The process is shown in Figure 1, class vector and token vector are denoted by $T_{[CLS]}$ and T_i respectively. The $T_{[CLS]}$ vector or average value of other token vectors output by BERT will be input to the classification layer to complete the paper classification.

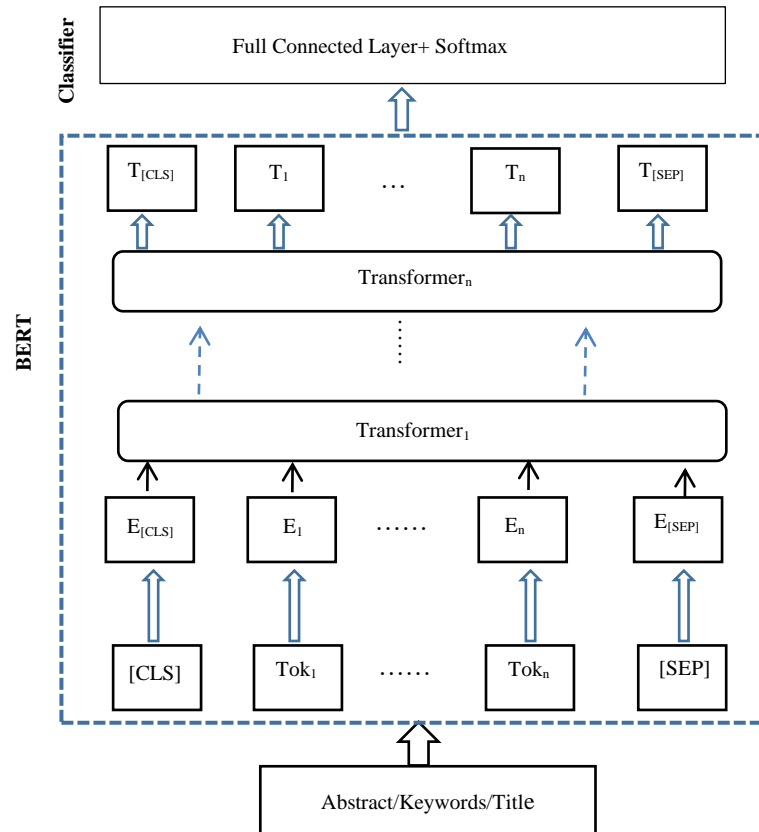


Figure 1. Paper classification by using BERT model

3.3. Classification modeling

CNN and RNN models require preprocessing the text sequences before classification. After the text sequences were tokenized, a vocabulary was generated. Word embedding vectors can be randomly generated, but using public pre-trained word embedding vectors can achieve better classification results. In our study, the Chinese pre-trained word vectors were used, which were obtained by training the Word2vec model on the Baidu Encyclopedia corpus (Word+Character, 300d) [27].

3.4. The experiment setup

To train and test the models, we extracted 1,000 records per discipline from the CSL dataset to construct the experimental dataset, including abstract, keywords, title, and category fields. The experimental dataset is divided into a training set, validation set, and test set according to the ratio of 7:2:1. The hardware configuration of the experimental platform and model development framework are shown in Table 3. We used four methods to compare and analyze the performance of paper classification, including single sentence input and concatenation method, sentence pair input method, ensemble learning method, and combination of BERT and CNN/RNN models. The definitions of these methods and the experimental purposes are shown in Table 4.

Table 3. Experimental environment parameters

Parameter	Value
CPU	Intel(R) Xeon(R) Silver 4216, 2.8G
RAM	128G
GPU	RTX 3090
Model development framework	PyTorch 1.13.1

Table 4. Different methods for paper classification

No.	Method	Description	Purpose
1	Single sentence input and concatenation method	Abstract, keywords, and title are regarded as a single sentence and input into the BERT model for classification. Use different methods to concatenate abstract, keywords, and title before entering the model for classification.	Compare the importance of the Abstract, keywords, and title in paper classification, and analyze the impact of connecting multiple fields as input to the model on classification performance.
2	Sentence pairs input method	Combine the Abstract, keywords, and title into sentence pairs in different ways, and then input them into the BERT model for classification.	Analyze the impact of different combination methods on classification performance in the sentence pair mode.
3	Ensemble learning method	The BERT model is used to classify Abstracts, keywords, and titles respectively, and the classification results of the three models are integrated to obtain the final classification result.	Compare the impact of integrating the classification results of the three models in different ways on the final classification results.
4	Combination of BERT and CNN/LSTM model	Use the BERT model to obtain the sentence representation feature vector of the abstract, and then input it into the CNN/LSTM model for classification.	Compare the paper classification performance when combining BERT with other models.

In order to evaluate the performance of the model in paper classification, we introduce 4 metrics, including accuracy, weighted average precision, weighted average recall, and weighted average F1-score, which are defined in (1) to (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 - \text{score} = \frac{2TP}{2TP+FP+FN} \quad (4)$$

Among them, true positive (TP) represents the number of correctly predicted positive samples, false positive (FP) represents the number of incorrectly predicted positive samples, true negative (TN) represents the number of correctly predicted negative samples and false negative (FN) represents the number of negative samples incorrectly predicted. Since paper classification is a multi-classification problem, we can calculate the precision, recall, and F1-score values of each category first, and then the weighted average precision, weighted average recall, and weighted average F1-score metrics are used to measure the overall classification performance for all categories. Taking the weighted average F1-score as an example, its calculation formula is in (5).

$$\text{Weighted Average } F_1 - \text{score} = F_{1_{\text{class1}}}W_1 + F_{1_{\text{class2}}}W_2 + \dots + F_{1_{\text{classN}}}W_N \quad (5)$$

$F_{1_{\text{classN}}}$ represents F1-score of class N, W_N represents the weight of each class, which is determined by the ratio of the number of samples in classN to the total number of samples.

4. RESULTS AND DISCUSSION

This study investigates the different effects of titles, keywords, and abstract fields on paper classification. Although earlier studies explored the effects of individual fields, they did not explicitly address the effects of different field combinations. We have tested three different combination methods in BERT model.

4.1. Single sentence input and concatenation method

The input of text sequence for the BERT model can be in the form of a single sentence or a pair of sentences. To evaluate the impact of the abstract, keywords, and title of the paper on the classification effect of the paper, the abstract, keywords, and title are treated as independent sentences, and then use the BERT/CNN/LSTM model to calculate the classification accuracy in the experimental data set. In addition, different concatenated forms between abstracts, keywords, and titles, including abstract+title, abstract+keywords, and abstract+keywords+title, are treated as sentences to evaluate the impact of different types of information combinations on classification performance. The experimental results are shown in Table 5. It can be seen that using abstract, keywords, and title alone as the input data of the BERT model, inputting abstract can achieve the highest classification accuracy, which is significantly higher than keywords or title. Compared with CNN and LSTM models, the BERT model can achieve higher classification performance. For sentences containing more than two elements from abstract, keywords, and title, the methods that combining abstracts with keywords, or combining abstracts with titles, can slightly improve classification accuracy over using only one element. However, the classification effect of combining abstract, keywords and titles is not as good as the former. It shows that when more sentence information is input, more noise data may also be imported.

Table 5. Single sentence & concatenation method classification

Model	Input data (Single sentence)	Accuracy	Weighted average precision	Weighted average recall	Weighted average F1 score
BERT	Abstract	0.8690	0.8689	0.8688	0.8673
CNN	Abstract	0.8007	0.7975	0.8007	0.7970
LSTM	Abstract	0.7939	0.7908	0.7939	0.7899
BERT	Title	0.8214	0.8181	0.8214	0.8182
BERT	Keywords	0.8200	0.8200	0.8201	0.8185
BERT	Abstract+title	0.8720	0.8723	0.8725	0.8709
BERT	Abstract+keywords	0.8707	0.8701	0.8707	0.8693
BERT	Abstract+keywords+title	0.8680	0.8679	0.8681	0.8664

4.2. Sentence pairs input method

Inputting the abstract, keywords and title of the paper into the BERT model in the form of sentence pairs actually allows the model to learn the relationship between the two sentences. The category to which the paper belongs can be regarded as a relationship. Classification of papers is achieved by learning the implicit association information of Abstract, keywords and title. The three types of sentence pairs <Abstract, Title>, <Abstract, Keywords>, <Abstract, Title+keywords> will be used as input to the BERT model, and their classification performance will be evaluated. Among them, <Abstract, Title+keywords> means that title and keywords are first concatenated into a sentence, and then combined with abstract to form a sentence pair. According to the experimental results, the classification effect of sentence pairs using the <Abstract, Title+keywords> method in the BERT model is slightly better than the other two methods. The experimental results are shown in Table 6.

Table 6. Sentence pair classification result

Model	Input Data (Sentence Pair)		Accuracy	Weighted average precision	Weighted average recall	Weighted average F1 score
	Sentence A	Sentence B				
BERT	Abstract	Title	0.8811	0.8806	0.8811	0.8788
BERT	Abstract	Keywords	0.8860	0.8846	0.8858	0.8841
BERT	Abstract	Title + Keywords	0.8880	0.8867	0.8875	0.8865

4.3. Ensemble learning method

It can be seen from the previous experimental results that the BERT model is used to classify the abstract, keywords and title of the paper, and the accuracy is 86.9%, 82.0% and 82.1% respectively. Although the sequence length of abstracts far exceeds keywords and titles, the latter two types of text sequences still contain important information that can distinguish different categories. Therefore, integrating the output results of the BERT models after processing the abstract, keywords and title respectively may improve the final paper classification accuracy. The [CLS] token output vectors obtained after the BERT model processes abstract, keywords, and title respectively are represented as BERT_Abstract, BERT_Keyword and BERT_Title in turn. The [CLS] output vectors of different types are summed and then input into the classification layer to complete

the classification of the paper. Experimental results show that the classification performance of summing BERT_Abstract, BERT_Keyword and BERT_Title as the input of the classifier is better than using other methods, such as summing BERT_Abstract and BERT_Keyword, summing BERT_Abstract and BERT_Title, or using BERT_Abstract alone. The experimental results are shown in Table 7.

Table 7. Ensemble learning method classification result

Model	Input data (Single sentence)	Accuracy	Weighted average precision	Weighted average recall	Weighted average F1 score
BERT	BERT_Abstract	0.8690	0.8689	0.8688	0.8673
BERT	BERT_Abstract+BERT_Keywords	0.8730	0.8766	0.8730	0.8724
BERT	BERT_Abstract+BERT_Title	0.8689	0.8694	0.8689	0.8684
BERT	BERT_Abstract+BERT_Keywords+BERT_Title	0.8740	0.8743	0.8745	0.8719

4.4. Combination of BERT and CNN/LSTM models

Since the BERT model is good at acquiring semantic information of text sequences, the text feature vectors it outputs can be used as input feature vectors for other models, so that the advantages of various models can be comprehensively utilized to improve the performance of paper classification. We used the paper abstract as input data, obtained the token's feature representation vector through the BERT model, and then input it into the CNN, LSTM, and RCNN models for classification. The CNN model used 256 convolution kernels with sizes of 1, 2 and 3, maximum pooling method is used to reduce the dimension of the output features. The RCNN model will use the formula in (6).

$$\text{Output} = \text{MaxPool}(\text{LSTM}(\text{BERT_output}) + \text{BERT_output}) \quad (6)$$

The experimental results are shown in Table 8. It can be seen that the classification performance of the BERT+CNN and BERT+RCNN models is improved compared to the BERT model alone. This shows that the CNN model's ability to obtain local features of text sequences can improve the classification performance of the BERT model.

Table 8. Combination of BERT and other models

Model	Input data (Single sentence)	Accuracy	Weighted average precision	Weighted average recall	Weighted average F1 score
BERT	Abstract	0.8690	0.8689	0.8688	0.8673
BERT+CNN	Abstract	0.8739	0.8736	0.8739	0.8725
BERT+LSTM	Abstract	0.8681	0.8700	0.8681	0.8683
BERT+RCNN	Abstract	0.8737	0.8743	0.8737	0.8721

5. CONCLUSION





In order to study how to effectively utilize abstract, keyword, and title information to achieve automatic classification of Chinese papers, multiple input data processing methods, and multiple deep learning models were applied to the experimental data set. Finally, we can draw the following conclusions: i) the effect of using the BERT model alone to classify papers is significantly better than using the CNN or LSTM model alone; ii) in paper classification, using abstract alone as the input data of the BERT model, the classification accuracy of the model is significantly better than using keywords or title alone as input data. After connecting abstract, title and keywords in different ways as input data, the classification accuracy is not significantly improved; iii) combine abstract, title, and keywords in different ways into sentence pairs as input data for the BERT model. The classification performance is significantly improved compared to the single sentence input form; iv) the BERT model is used to process abstract, title, and keywords respectively, and the three output results are summed in different combinations and used as input data for the classification layer. The experimental results show that the classification effect is slightly better by taking the sum of the output results of the three BERT models than using one BERT model alone; v) treat the BERT model as a text feature extractor, and the obtained text feature vectors are then input into CNN, RNN, or other models for secondary processing, to comprehensively apply the capabilities of different types of deep learning models and extract more effective classification information. Experimental results show that the combination of BERT and CNN can enable the model to achieve better classification performance than other combination methods. However, compared to using the BERT model alone for classification, the performance improvement is not obvious; and vi) the experimental dataset itself also has factors that affect classification performance. Currently, training

samples are assigned only one category value. Some papers belong to interdisciplinary and there will be scope overlap between categories. Therefore, hierarchical and weighted multi-category paper classification is more promising. In the next step of research, we will also try to use the citation information of the paper as a supplementary field to improve the effect of paper classification.





REFERENCES

- [1] Q. Li *et al.*, "A survey on text classification: from traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, 2022, doi: 10.1145/3495162.
- [2] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: from text to predictions," *Information*, vol. 13, no. 2, 2022, doi: 10.3390/info13020083.
- [3] T. Yue, Y. Li, X. Shi, J. Qin, Z. Fan, and Z. Hu, "PaperNet: A dataset and benchmark for fine-grained paper classification," *Applied Sciences*, vol. 12, no. 9, 2022, doi: 10.3390/app12094554.
- [4] A. Rajan and M. Manur, "Aspect based sentiment analysis using fine-tuned BERT model with deep context features," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1250–1261, 2024, doi: 10.11591/ijai.v13.i2.pp1250-1261.
- [5] M. Gao, T. Li, and P. Huang, "Text classification research based on improved word2vec and CNN," in *Service-Oriented Computing – ICSOC 2018 Workshops*, 2019, pp. 126–135, doi: 10.1007/978-3-030-17642-6_11.
- [6] S. Abdul-Rahman, M. F. A. M. Ali, A. A. Bakar, and S. Mutalib, "Enhancing churn forecasting with sentiment analysis of steam reviews," *Social Network Analysis and Mining*, vol. 14, no. 178, pp. 1–17, 2024, doi:10.1007/s13278-024-01337-3.
- [7] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," *Proceedings of the National Conference on Artificial Intelligence*, vol. 3, pp. 2267–2273, 2015, doi: 10.1609/aaai.v29i1.9513.
- [8] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," *arXiv-Computer Science*, pp. 1–10, 2015.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186.
- [10] I. Beltagy, K. Lo, and A. Cohan, "SCIBERT: A pretrained language model for scientific text," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 3615–3620, doi: 10.18653/v1/d19-1371.
- [11] Maryanto, Philips, and A. S. Girsang, "Hybrid model for extractive single document summarization: utilizing BERTopic and BERT model," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1723–1731, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1723-1731.
- [12] D. Wei and Z. Jie, "Scientific literature classification research based on the density distribution of OCSVM," *Information Engineering*, vol. 4, no. 3, pp. 67–72, 2018, doi: 10.3772/j.issn.2095-915x.2018.03.009.
- [13] M. I. Elias, Y. Mahmud, S. Mutalib, S. N. K. Kamarudin, R. Maskat and S. A. Rahman, "Fake news prediction using hybrid model–systematic literature review," *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, pp. 281–286, 2023, doi: 10.1109/AiDAS60501.2023.10284628.
- [14] I. Jaya, I. Aulia, S. M. Hardi, J. T. Tarigan, M. S. Lydia, and Caroline, "Scientific documents classification using support vector machine algorithm," *Journal of Physics: Conference Series*, vol. 1235, no. 1, pp. 1–6, May 2019, doi: 10.1088/1742-6596/1235/1/012082.
- [15] Y. Xiaohua and G. Haiyun, "Improved Bayesian algorithm based automatic classification method for bibliography," *Computer Science*, vol. 45, no. 8, pp. 203–207, 2018, doi: 10.11896/j.issn.1002-137X.2018.08.036.
- [16] M. Ech-Chouyyek, H. Omara, and M. Lazaar, "Scientific paper classification using convolutional neural networks," in *ACM International Conference Proceeding Series*, 2019, pp. 1–6, doi: 10.1145/3372938.3372951.
- [17] G. A. Burns, X. Li, and N. Peng, "Building deep learning models for evidence classification from the open access biomedical literature," *Database*, vol. 2019, no. 1, 2019, doi: 10.1093/database/baz034.
- [18] M. Samami and E. M. Soure, "Binary classification of Lupus scientific articles applying deep ensemble model on text data," in *2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*, 2019, pp. 12–17, doi: 10.1109/ICDIPC.2019.8723787.
- [19] A. Bogdanchikov, D. Ayazbayev, and I. Varlamis, "Classification of scientific documents in the kazakh language using deep neural networks and a fusion of images and text," *Big Data and Cognitive Computing*, vol. 6, no. 4, pp. 1–12, Oct. 2022, doi: 10.3390/bdcc6040123.
- [20] R. Sarasu, K. K. Thyagarajan, and N. R. Shanker, "SF-CNN: Deep text classification and retrieval for text documents," *Intelligent Automation and Soft Computing*, vol. 35, no. 2, pp. 1799–1813, 2023, doi: 10.32604/iasc.2023.027429.
- [21] S. Lili, J. Peng, and W. Jing, "A study on the automatic classification of chinese literature in periodicals based on BERT model," *Library Journal*, vol. 41, no. 5, 2022, doi: 10.13663/j.cnki.lj.2022.05.014.
- [22] Z. Yang, Z. Zhixiong, L. Huan, and D. Liangping, "Classification of chinese medical literature with bert model," *Data Analysis and Knowledge Discovery*, vol. 4, no. 8, pp. 41–49, 2020, doi: 10.11925/infotech.2096-3467.2019.1238.
- [23] X. Hongling, F. Guohe, and H. Weilin, "Research on semantic classification of scientific and technical literature based on deep learning," *Information studies: Theory & Application*, vol. 41, no. 11, pp. 149–154, 2018, doi: 10.16353/j.cnki.1000-7490.2018.11.027.
- [24] K. Jie, "Research on automatic literature classification system based on deep learning and chinese library classification," *New Century Library*, vol. 5, pp. 51–56, 2021, doi: 10.16810/j.cnki.1672-514X.2021.05.009.
- [25] Y. Zhang *et al.*, "Weakly supervised multi-label classification of full-text scientific papers," in *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, California, Aug. 2023, pp. 3458–3469, doi: 10.1145/3580305.3599544.
- [26] Y. Li *et al.*, "CSL: A large-scale chinese scientific literature dataset," *Proceedings - International Conference on Computational Linguistics, COLING*, vol. 29, no. 1, pp. 3917–3923, 2022.
- [27] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on chinese morphological and semantic relations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, vol. 2, Jul. 2018, pp. 138–143, doi: 10.18653/v1/P18-2023.





BIOGRAPHIES OF AUTHORS

Xin Luo     is pursuing in Computer Science in School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. His current research interests are deep learning and natural language process. He can be contacted at email: 2022201126@isiswa.uitm.edu.my.



Sofianita Mutalib     is currently the associate professor in School of Computing Sciences, College of Computing, Informatics and Mathematics Universiti Teknologi MARA, (UiTM) Shah Alam. She teaches bachelor and postgraduate courses related to intelligent systems such as intelligent system development, data mining, and final project. Her primary research interests involve intelligent systems, data mining as well as machine learning and also data science. She can be contacted at email: sofianita@uitm.edu.my.



Syarifah Ruzaini Syed Aris     is currently a senior lecturer in School of Computing Sciences, College of Computing, Informatics and Mathematics Universiti Teknologi MARA (UiTM), Shah Alam. Her primary research interests involve strategic management information systems and business intelligence. She can be contacted at email: ruzaini@uitm.edu.my.