Research on Automatic Classification Based on Academic Text Corpus

Chuan Jiang, Zhixiao Zhao, Na Wu, Litao Lin
School of Information Management, Nanjing Agricultural University, Nanjing 210095, China.

Abstract: Recognizing the discipline category of the abstract text is of great significance for automatic text recommendation and knowledge mining. Therefore, this study obtained the abstract text of social science and natural science in the Web of Science 2010-2020, and used the machine learning model SVM and deep learning model TextCNN and SCI-BERT models constructed a discipline classification model. It was found that the SCI-BERT model had the best performance. The precision, recall, and F1 were 86.54%, 86.89%, and 86.71%, respectively, and the F1 is 6.61% and 4.05% higher than SVM and TextCNN. The construction of this model can effectively identify the discipline categories of abstracts, and provide effective support for automatic indexing of subjects.
Keywords: Deep Learning; SCI-BERT; Academic Literature; Automatic Indexing

1. Introduction

As one of the basic research tasks in knowledge mining, automatic text classification can not only effectively distinguish the categories contained in the text, but also play an important role in promoting subsequent automatic recommendation, automatic text summarization and text structure knowledge mining. The research of automatic classification on general text has achieved great success, and has the value of engineering application to a certain extent. At the same time, with the rapid development of deep learning technology, the classification effect on general text is more prominent. However, the research on domain text needs to be strengthened and deepened. On the one hand, the scale of corpus used in automatic classification of domain text is relatively small on the whole, and on the other hand, the technology used needs to be improved.

In recent years, with the convenience of obtaining corpus in the academic field, the research on the exploration of academic text corpus based on deep learning has increased more. However, by combing the relevant research, it is found that the research on automatic classification in the academic field mainly focuses on several disciplines or based on small-scale corpus. From the perspective of the whole academic inquiry, it is of special value and significance to classify and explore the whole discipline based on the large corpus. On the basis of the above, based on the corpus of Web of Science 2010-2020, this paper constructs the corresponding automatic classification model through traditional machine learning and deep learning. On the one hand, this study helps to classify a paper from the whole subject, on the other hand, it also verifies the performance of deep learning and traditional machine learning from the perspective of model.

2. Literature review

SVM is a classic classification algorithm. Since it was proposed, many scholars have carried out extensive research on SVM-based feature selection methods and application scenarios (Chandra and Bedi, 2021). Neumann et al.(2006) presented four novel continuous feature selection approaches that directly minimize classifier performance, using both linear and nonlinear Support Vector Machine classifiers. The authors' methods include additional regularization and embedded nonlinear feature selection, and they apply difference of convex functions programming to solve their optimization problems. The authors demonstrate the effectiveness of their proposed method through experiments on organ image classification. Bazi Y et al.(2008) studied how to improve the performance of SVM when it is applied to the classification task of hyperspectral imagery, and proposed a hyperspectral image classification system based on SVM. The system allows detection of optimal discriminative features and automatic estimation of optimal SVM parameters through a genetic optimization framework. Mathur et al.(2008) explores the potential of support vector machines for supervised

classification analyses in remote sensing. The results show that the one-shot multiclass classification approach is more accurate and efficient than the traditional binary SVM-based approaches. Pal et al.(2010) explored the accuracy of SVM classification and its sensitivity to the number of features used in the task of the classification of hyperspectral data. Rodrigo et al.(2013)compared the effectiveness of Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in sentiment analysis. According to the experiment, the performance of SVM and ANN in sentiment analysis depends on the dataset and the level of data imbalance. In terms of classification accuracy on the benchmark dataset of Movies reviews, ANN outperformed SVM significantly, even in the context of unbalanced data. However, on datasets with more noisy terms, such as Books, GPS, and Cameras, the performance of ANN tended to decrease below the performance of SVM as the data imbalance increased. In terms of efficiency, the study found that SVM had a high computational cost at the running time, while ANN had a high computational cost at the training time. Ebrahimi et al.(2017) used support SVM method to detect thrips on the crop canopy images. In this study, the image processing technology was combined with the branch SVM method, and the target thrips could be detected accurately on the basis of selecting the appropriate area and color index, and the error was less than 2.5%.

SVM offers a principled approach to machine learning problems because of its mathematical foundation in statistical learning theory (2015). It can be seen from the above research that SVM can provide great help in both natural science and social science research. In the past few decades, various variants of SVM have emerged and have been applied to the classification tasks of text and images in different fields. Even in the era of neural network models, SVM has outstanding performance advantages in some tasks.

Liu and Guo (2019) proposed a new AC-BiLSTM model, for the high dimensionality, sparsity, and semantic complexity of data faced in text classification problems, which contains BiLSTM, attention mechanism, and the convolutional layer to capture the local features of a phrase and the global semantics. Yue and Li (2020) proposed a text classification method that incorporates Word2vec, BiLSTM, and CNN, in order to solve the problem that short text classification algorithms are prone to find errors when performing sentiment classification, and the study showed that this hybrid network model performs better than a single-structure neural network model in short texts. Deng et al. (2021)proposed ABLG-CNN model, a text classification model that fuses CNN with gating mechanism, which is based on BiLSTM and can capture semantic and local phrase features of text context. Wu et al. (2021)proposed a multi-category text classification model WTCM based on weighted Word2vec, BiLSTM and attention mechanism, which was shown to be effective in solving the problems of high latitude of text vector representations and weak ability of semantic feature information extraction in traditional multi-category text categorization algorithms. Enamoto et al. (2022)proposed a shallow network with a BiLSTM layer and an Attention layer, in order to solve the problem of BiLSTM in Portuguese legal text classification, and the study showed that combining the BiLSTM layer and the Attention layer helps to capture the contextual information in long judicial texts. Guo et al. (2022)proposed a text classification method based on BiLSTM and Attention mechanism for the problem of difficult to accurately categorize and manage Internet news, and the study showed that the method can effectively improve the classification effect of Chinese long news text. Ruan et al. (2022)proposed an ATT-CN-BILSTM Chinese news classification model by combining CNN model and BiLSTM model, which improves the feature extraction process of CNN and BiLSTM by using attention mechanism. Xue et al. (2022) proposed an improved Mutual Graph Convolution Networks (IMGCN) model, which introduces a semantic dictionary (WordNet), contextual dependencies, BERT, and bi-directional long and short-term memory network (BiLSTM) to solve the problem of Graph Convolutional Neural Networks (GCN) cannot fully utilize context-dependent information and is not good at capturing local information in text classification. Xu et al. (2022)proposed a multi-model structure based on Bert-CNN-BiLSTM for the overfitting phenomenon of the BERT pre-training model in training a small-sample Chinese text classification dataset, which was shown to be effective in extracting the feature information in the text.

The above studies cover a variety of BiLSTM-based neural network architectures for text classification tasks. By incorporating the attention mechanism, convolutional layers, and different neural network models, such as Word2vec, BERT, CNN, etc., these methods effectively capture the local and global features of long and short texts by incorporating attention mechanisms, convolutional layers and different neural network models, such as Word2vec, BERT, CNN, etc. It performs well in solving the problems of long-term dependence, semantic complexity and high dimensionality of data, and improves the text classification effect of BiLSTM model in dealing with long text, short text, small sample and other types of data. These studies confirm that the hybrid network

structure performs better than the single network structure in text classification tasks, and provides innovative ideas and methods for solving problems in practical applications.

Sun et al. (2019) comprehensively explored the performance of BERT under different fine-tuning strategies on various classification datasets, such as sentiment classification and topic recognition. They found that further pre-training can significantly improve performance, and both single or multi-task fine-tuning has a positive impact on the classification results. Li et al. (2019) fed the original BERT embedding to CNN and Bi-LSTM to further extract local and global features, and combined them through fully connected layer. Their method surpassed the previous model in the job recruitment text classification task. Lu et al. (2020) proposed the VGCN-BERT text classification model, which combined BERT and Vocabulary Graph Convolutional Network (VGCN) to enhance the learning capability of global vocabulary information. It surpassed the performance of single model on five classification datasets, including sentiment analysis and hate detection. Maheshwari et al. (2021) classified the citation purpose and importance of academic text citation contexts based on the BERT framework. BERT and SciBERT achieved the best results respectively, and linear classifier was better than Bi-LSTM. Mondal (2021) proposed the BBAEG (Biomedical BERT-based Adversarial Example Generation) attack algorithm for biomedical text classification. This method enhanced the performance of BERT-MLM in entity and synonym replacement, and introduced multiple mechanisms for entity modification and generating high-quality adversarial samples. Pujari et al. (2021) designed a multi-label academic text classification model. They used several multi-layer perceptrons to predict the probability of each category label for metadata embedded with SciBERT, and verified the performance on BioCreative VII task. Khadhraoui et al. (2022) constructed the Cov-Dat-20 abstract dataset containing four COVID-19 related topics. They further pre-trained BERT on this dataset and got the CovBERT model specialized for COVID-19 related scientific text. Smirnova and Mayr (2023) compared the named entity recognition (NER) performance of different Flair embedding models on corpora of varying scales. They found that adding RoBERTa embedding did not improve performance, and when the corpus size increased to a certain extent, the NER performance dropped. Mou et al. (2023) proposed a unified pre-training architecture (UPPAM) for language-based political actor modeling, which introduced new mapping relations, new pre-training tasks, and multi-granularity actor representation models during pre-training. It achieved better performance in several downstream political text classification tasks. KafiKang and Hendawi (2023) combined Relation BioBERT (R-BioBERT) embedding with Bi-LSTM classifier for drug-drug interaction (DDI) extraction. Their method surpassed baseline performance on SemEval 2013, TAC 2018 and TAC 2019 datasets.

The above-mentioned studies indicate that current academic text classification tasks mainly adopt the methods based on BERT embedding. Researchers enhance BERT's classification performance on domain-specific texts by employing strategies such as continuing pre-training, modifying self-supervised learning tasks, and fine-tuning. Particularly, the incorporation of neural networks like CNN and GCN further extracts text features, enabling more effective learning of both local and global information. Additionally, besides common multi-class text classification tasks, there has been some exploration into the more complex multi-label classification.

3. Data source

The corpus of abstracts used in this study was obtained from the WOS (Web of Science) database, with abstract data selected from the years 2010-2020. Firstly, all the title information of the WOS database from 2010-2020 was obtained, after which the missing data of year, abstract, and WOS categories were filtered to obtain the preliminary literature categories and abstract data. In terms of abstract text category determination, the number of subcategories provided in the WOS database is too large, while a document may have multiple subject categories, which is not conducive to subsequent text classification studies targeting abstracts. Therefore, this study adopts the five categories set by the WOS database as the reference for abstract text categories.Specifically, according to the classification system provided by the WOS database, the category information in the title information of the WOS database is divided into five categories. In this process, for documents belonging to multiple WOS categories, after statistical investigation, multiple WOS categories belonging to the same document often belong to the same major category, so the first WOS category belonging to the document is selected as the reference for category classification. After screening and extracting the title information of WOS database from 2010 to 2020, a total of 9560290 abstracts were obtained, and the overall statistics of abstract data are shown in Table 1.

Table 1 Statistical information of WOS abstract corpus

Basic statistical information	Values

Total file size	11.55 GB
Total number of texts	9,560,290
Total number of words	1,780,769,324
Average sentence length of abstracts	186
Maximum sentence length of abstracts	3059
Minimum sentence length of abstracts	50

Table 2 demonstrates the distribution of categories of summary data from the wos database used in this study. It can be seen that the Life Science & Biomedicine discipline contains the largest number of sub-disciplines as well as the number of documents, while the Arts & Humanities discipline contains only 0.21% of the total number of abstracts. On the one hand, Arts & Humanities has the least number of subdisciplines in the classification system used in this paper, and on the other hand, in the process of data extraction and classification, there is a part of the wos classification name does not correspond to the name of the category in the classification system, which leads to the small proportion of abstracts in Arts & Humanities. At the same time, the number of abstracts of Social Sciences is much less than Physical Sciences and Technology when they have similar number of subdisciplines. WOS database, as a more comprehensive database, shows that Humanities & Social Sciences research is weaker than natural sciences.

Table 2 The distribution of WOS abstracts in accordance with discipline

Categories	**Number of sub-disciplines**	**Total with**	**%**
Life Sciences & Biomedicine	64	4,097,398	42.86%
Physical Sciences	16	2,804,019	29.33%
Technology	20	2,018,146	21.11%
Social Sciences	19	620,369	6.49%
Arts & Humanities	7	20,358	0.21%
Total	**126**	**9560290**	

4. Methods

4.1 Support Vector Machine (SVM)

As shown in Figure 1, the SVM model is a classification model based on statistical theory. The basic idea is to construct a hyperplane as a decision plane to maximise the distance between positive and negative modes. This model has wide applications in terms of text classification. This research applied it to the task of the identification of abstract category. Firstly, we transformed the abstract sentences into a term frequency–inverse document frequency (TF-IDF) vector. Then, by nonlinear mapping, the SVM model mapped the vectorised abstract sentences TF-IDF vector to a high-dimensional feature space, converted the nonlinear classable problem in the original sample space into a linear classable problem in this feature space, and used kernel functions to avoid the curse of dimensionality and reduce computational complexity. Applying a margin maximisation learning strategy to adjust the model parameters enables the classification of abstract category.

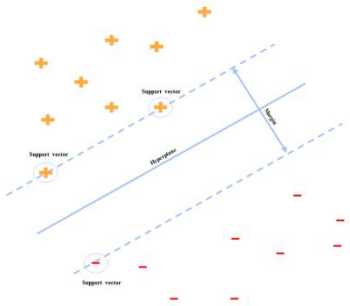

Figure 1. discipline classification model based on SVM

4.2 Text convolutional neural network (TextCNN)

As shown in Figure 2, the TextCNN model (Kim 2014) is a text classification model based on CNNs. First, through a word

embedding layer, each word of the abstract is mapped into a word embedding representation to form a word embedding matrix of abstract, and then through a convolution layer, with different filters, the abstract word embedding matrix can be scanned and calculated in the manner of a sliding window, which is similar to extracting N-gram, obtaining abstract convolutional semantic vectors. Then, through a Pooling layer, the most effective features in the abstract convolution features for abstract category classification can be extracted using Max Pooling, thereby obtaining the abstract Pooling semantic vectors. Finally, using the Softmax layer enables classifying and predicting the semantic vector features of abstract Pooling to obtain the abstract category and adjusting the TextCNN model parameters according to the prediction results.

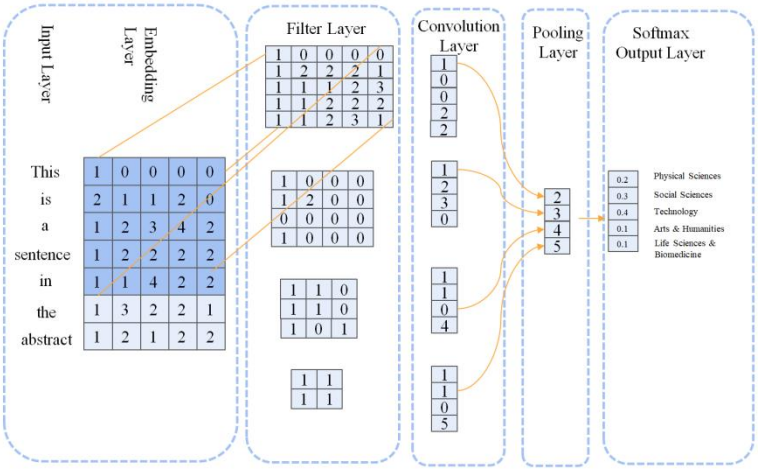

Figure 2. Discipline classification model based on TextCNN

4.3 Bidirectional Encoder Representation from Transformers (SCI-BERT)

As shown in Figure 3, BERT (Devlin et al. 2018) is a deep language representation model that is improved based on the bidirectional language model. SCI-BERT is a pre-trained model based on BERT using academic literature. It is completely based on the self-attention mechanism transformer structure to model the sentences in abstracts. The advantage of SCI-BERT compared to other neural network models is that a large-scale unsupervised corpus is adopted for pre-training. When applying the abstract category classification task, the initial parameters of the entire training process are originated from the pre-training model. First, abstract word embedding, sentence embedding, and position embedding vectors are added, and then, through self-attention multi-layers, the semantic vectors of abstract are made available, and finally through the Softmax layer, the category of abstract is classified and predicted. Model parameters can be fine-tuned according to the prediction results.

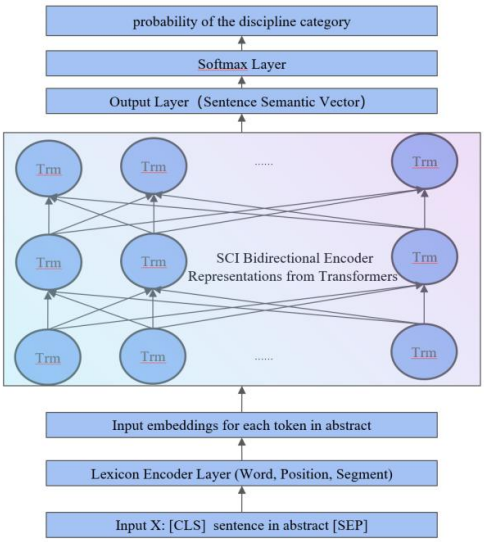

Figure 3. discipline classification model based on SCI-BERT

5. Experiment

5.1 Model parameters and experiment environments

The SVM was implemented using the sklearn library in Python, where the kernel function is Radial Basis Function (RBF); the penalty coefficient was set to C=2.0, gamma=0.5, and text representation used TF-IDF.

TextCNN and SCI-BERT were the deep learning models for text classification. The main structure of the TextCNN model was convolution neural network, the dimension of word embedding was set to 128, the number of filters was 200, the filter size was 3*3, 4*4, 5*5. To prevent the training data from overfitting, the dropout rate was set to 0.5 and the epoch was set to 20. The batch size was 512. The main structure of the SCI-BERT model was Transformer; it uses transfer learning to change the output layer of pre-trained SCI-BERT model, which is used for academic abstract subject classification task in the field of social sciences and natural sciences, where the number of hidden unit was set to 768, the number of self-attention was 12, warm-up proportion was set to 0.1, learning rate was 2.0E-5, the batch size was 16, the maximum sequence length was 512, and the epoch was set to 3.

Since the neural network involves massive matrix calculations during the training process, in order to accelerate the training speed, this research used NVIDIA Tesla P40 GPU to train the neural network. The main parameters of the testing machine were: CPU: 48 Intel (R) Xeon (R) CPU E5-2650 v4 @ 2.20GHz; Memory: 256GB; GPU: 6 NVIDIA Tesla P40 memory: 24GB; Operating system: CentOS 3.10.0.

5.2 Experimental results

In order to test the accuracy of the model, we divided the training data in the form of 8:2, that is, 80% of the data is the training set, and 20% of the data is the test set. Due to the large amount of data, we sampled the data, each discipline according to the standard of 50,000 abstracts, if it is less than 50,000, all will be taken out, the overall abstract data is shown in Table 3

Table 3 Distribution of sampling data

Discipline	number of abstracts
Physical Sciences	50000
Social Sciences	50000
Technology	50000
Arts & Humanities	20358
Life Sciences & Biomedicine	50000

As shown in Table 4, we use the macro average method to calculate the precision, recall, and F1. Among them, SCI-BERT has the highest performance, with precision rate, recall rate, and f1 reaching 86.54%, 86.89%, and 86.71% respectively. The precision is 3.76% and 6.11% higher than SVM and TextCNN, the recall is 4.33% and 6.75% higher than SVM and TextCNN, and the f1 is 4.05% and 6.61% higher than SVM and TextCNN. The reason for this is that SCI-BERT has been pre-trained on a large-scale unsupervised paper corpus, and uses a Transformer architecture with a self-attention mechanism, which works well.

Table 4 Classification performance comparison of different models

	SVM	TextCNN	SCI-BERT
macro-avg precision	82.78%	80.43%	86.54%
macro-avg recall	82.56%	80.14%	86.89%
macro-avg f1	82.66%	80.10%	86.71%

6. Conclusion

This study first obtained the WOS 2010-2020 abstract texts of natural science and social science, and built a basic corpus with a total of 9,560,290 abstract texts. Then classify its subjects, and construct 5 categories in total, namely Physical Sciences, Social Sciences, Technology, Arts & Humanities, Life Sciences & Biomedicine. Then, the subject classification model was constructed by using the machine learning SVM model and the deep learning models TextCNN and SCI-BERT. It was found that the SCI-BERT model had the best performance, and the precision, recall, and f1 were 86.54%, 86.89%, and 86.71% respectively, the f1 is 6.61% and

4.05% higher than the TextCNN and SVM models, respectively. In the future, we will build support for extracting semantic knowledge in abstracts, and provide support for automatic indexing of academic literature knowledge.

References

[1] Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector machines for classification. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, 39-66.

[2] Bazi, Y., & Melgani, F. (2006). Toward an optimal SVM classification system for hyperspectral remote sensing images. IEEE Transactions on geoscience and remote sensing, 44(11), 3374-3385.

[3] Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification. *International Journal of Information Technology*, 13, 1-11.

[4] Deng, J., Cheng, L., & Wang, Z. (2021). Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. *Computer Speech & Language*, 68, 101182.

[5] Ebrahimi, M. A., Khoshtaghaza, M. H., Minaei, S., & Jamshidi, B. (2017). Vision-based pest detection based on SVM classification method. *Computers and Electronics in Agriculture*, 137, 52-58.

[6] Enamoto, L., Santos, A. R., Maia, R., Weigang, L., & Filho, G. P. R. (2022). Multi-label legal text classification with BiLSTM and attention. *International Journal of Computer Applications in Technology*, 68(4), 369-378.

[7] Guo, H., Zhang, J., & Xiao, L. (2022, August). A news text classification method based on the BiLSTM-Attention. *In 2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)* (pp. 468-472). IEEE.

[8] KafiKang, M., & Hendawi, A. (2023). Drug-Drug Interaction Extraction from Biomedical Text using Relation BioBERT with BLSTM. *Machine Learning and Knowledge Extraction*, 5(2), 669-683.

[9] Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. (2022). Survey of BERT-base models for scientific text classification: COVID-19 case study. *Applied Sciences*, 12(6), 2891.

[10] Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019, December). The automatic text classification method based on bert and feature union. *In 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)* ,774-777.

[11] Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337, 325-338.

[12] Lu, Z., Du, P., & Nie, J. Y. (2020). VGCN-BERT: augmenting BERT with graph embedding for text classification. *In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020*, 14–17.

[13] Maheshwari, H., Singh, B., & Varma, V. (2021, June). Scibert sentence representation for citation context classification. *In Proceedings of the Second Workshop on Scholarly Document Processing* ,130-133.

[14] Mathur, A., & Foody, G. M. (2008). Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2), 241-245.

[15] Mondal, I. (2021). BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. *arXiv preprint arXiv*:2104.01782.

[16] Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621-633.

[17] Mou, X., Wei, Z., Zhang, Q., & Huang, X. J. (2023, July). UPPAM: A Unified Pre-training Architecture for Political Actor Modeling based on Language. I*n Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* , 11996-12012.

[18] Neumann, J., Schnörr, C., & Steidl, G. (2005). Combined SVM-based feature selection and classification. *Machine learning*, 61, 129-150.

[19] Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5), 2297-2307.

[20] Pujari, S. C., Tarsi, T., Strötgen, J., & Friedrich, A. Team RobertNLP at BioCreative VII LitCovid track: neural document classification using SciBERT. *In Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.

[21] Ruan, J., Caballero, J. M., & Juanatas, R. A. (2022, May). Chinese news text classification method based on attention mechanism. *In 2022 7th international conference on business and industrial research (ICBIR)* ,330-334.

[22] Smirnova, N., & Mayr, P. (2023). Embedding Models for Supervised Automatic Extraction and Classification of Named Entities in Scientific Acknowledgements. *arXiv preprint arXiv:2307*.13377.

[23] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. *In Chinese Computational Linguistics: 18th China National Conference, CCL 2019,* 194-206.

[24] Wu, H., He, Z., Zhang, W., Hu, Y., Wu, Y., & Yue, Y. (2021). Multi-class text classification model based on weighted word vector and bilstm-attention optimization. *In Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings,* 393-400.

[25] Xu, E., Qin, D., Huang, J., & Zhang, J. (2022, July). Multi text classification model based on bret-cnn-bilstm. *In 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI) ,184-189.*

[26] Xue, B., Zhu, C., Wang, X., & Zhu, W. (2022, March). The Study on the Text Classification Based on Graph Convolutional Network and BiLSTM. *In Proceedings of the 8th International Conference on Computing and Artificial Intelligence* ,323-331.

[27] Yue, W., & Li, L. (2020, December). Sentiment analysis using Word2vec-CNN-BiLSTM classification. *In 2020 seventh international conference on social networks analysis, management and security (SNAMS)*,1-5.