A Study on News Headline Classification Based on BERT Modeling

Yucheng Chen

Zhengzhou University of Light Industry,The College of Computer Science and Technology, Zhengzhou,Henan 450001,China
542107010330@zzuli.edu.cn

Abstract. News is an important way to understand the information of contemporary society, and it is necessary to quickly categorize and identify a large amount of news information.In this report, a classification task was performed on Chinese news headlines based on the Bidirectional Encoder Representations from Transformers (BERT) model. Deep learning model transformers are used to compare the differences between Bert model and traditional methods in text categorization. Training and tuning were performed on the collected and organized dataset.From the experimental results, the model has a better classification effect on news headline classification, reflecting the advantages and performance of Bert in Chinese news headline text classification. Meanwhile, the performance differences of Bert model under different learning rate parameters, number of learning rounds and different dataset annotation accuracy settings are analyzed.The results of the experiment were 0.7 accuracy for 10 rounds of learning with a learning rate parameter of 5e-6, and 0.6 accuracy for 20 rounds of learning with a learning rate parameter of 1e-5.The analysis concludes that under the same learning rate parameter, the learning accuracy tends to stabilize with the increase in the number of learning rounds; under the same number of learning rounds, the learning rate is too high or too low will affect the learning accuracy.

Keywords: Deep Learning, BERT , News Headline,Categorization.

1 Introduction

News is an important way for us to learn information about today's society, and in today's world where information is growing very rapidly, it is impossible for us to receive everything that is happening right now, so people often browse or receive news and information based on our different preferences for news categories.It is essential for us to be able to classify news in real time and receive personalized news recommendations [1]. For news organizations, it is important to be able to categorize a large amount of news in a timely manner and archive it quickly, so as to facilitate the analysis of public opinion and social hot spots [2].

Research has shown that text categorization is more widely used in natural language processing and information management. Bidirectional Encoder Representations from Transformers (BERT) is an improved model based on the Transformer encoder, which

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_35

belongs to the bi-directional language model and can make full use of contextual information.This model is suitable for text categorization natural language understanding tasks [3-5]. Traditional machine learning has limited representational capabilities in text categorization, with shallow semantic, structural and contextual understanding of text. Deep learning makes up for the shortcomings of traditional machine learning in text categorization and improves the ability to understand the context, but it also suffers from disadvantages such as poor model interpretability and difficulty in adjusting features.For text categorization traditional machine learning has limited characterization capability and shallow understanding of the semantics, structure and context of text. Deep learning makes up for many of the shortcomings of traditional machine learning [6].

Currently, research in the field of text categorization focuses mainly on text representation and classification models.Typically, text representation uses the vector space model or its variants to characterize each text as a vector, where the vector elements represent its semantic features [7-9]. However, the classification results for short texts are usually poor due to the fact that short texts are characterized by feature sparsity and Govette Call recruitment.To overcome this problem, various approaches have been proposed such as extracting association rules using external knowledge bases or internal semantic analysis to extend short text features [10]. However, the effectiveness of these methods is largely limited by the quality of the knowledge base and the complexity of the semantic analysis methods, and the difficulty of fully capturing the semantic information of short texts [11].

With the rapid development of deep learning technology, the field of text categorization has ushered in a new breakthrough.There is a mainstream trend towards text data processing methods based on BERT models, which are first pre-trained on large-scale corpora and can subsequently be fine-tuned for specific downstream tasks.Compared with traditional deep learning methods, this method shows superior performance and transferability [12]. This paper creates a four-classification dataset for Chinese news headlines, and use BERT to train its word vectors, extract the main features of the text, and then use the attention mechanism to weight and sum to adjust the proportion of the weights, which results in the final learning rate and classification accuracy.

2 Data and Methodology

2.1 Data Sources

The data used in this paper are Chinese news headlines, which are often the center of an article, so classifying news headlines can greatly reduce the content of the text and the amount of arithmetic, thus increasing the classification efficiency [13].

Data sources: Xinhua (www.xinhuanet.com), People's Daily (www.people.com.cn), China Daily (www.chinadaily.com.cn), China News (www.chinanews.com), Oriental (www.eastday.com) , Sohu News (news.sohu.com), Tencent News (news.qq.com), Sina News (news.sina.com.cn).

Among the many news sources , the above websites are more authoritative, which ensures the authenticity of the data in the dataset and reduces the workload of data

preprocessing. Using Python to crawl more than 2,000 news headlines in news websites into the dataset, eliminating duplicate text and classifying and labeling them according to uniform standards.

Write the text of the collected news headlines in a .txt file and categorize the news headlines into 0 social news, 1 political news, 2 economic news, and 3 scientific and technological news, and label all the data in the format of numerical markers + tab feeds.

The categorized collection was canonically labeled and then all aggregated into the total set, which was divided into a training set and a test set with a 7:3 ratio of the number of headline entries. After completing all labeling tasks browse through the check dataset again to confirm that the labeling is correct.

First of all, it is very important to establish a unified data classification standard system to classify and organize the data in a standardized way. Social news is news about social life, people's livelihood, social events and other aspects of the news, which involves the people's life, social customs, social issues and other content. Political news is news about national policies, government behavior , political figures' movements, political changes, etc., including domestic and foreign political developments and important political decisions. Economic news includes news about macroeconomics, industrial economy, financial market, enterprise management, economic policy, etc. It usually involves economic development, trade situation, financial data, etc. Science and technology news is about science and technology, innovation development, high-tech industry, Internet information technology and other aspects of the news, involving scientific and technological achievements, technology applications, science and technology policies and other content. News headlines are categorized according to this system of classification criteria above.

To ensure the objectivity of the data categorization, a team of annotators consisting of multiple journalism professionals was assembled to separate the data set into groups , ensuring that each group of data annotators was able to annotate the data independently.

Data are summarized at the end of independent annotation, discussed, analyzed and evaluated in a uniform manner, and the dataset is modified based on feedback.The dataset was again disrupted and redistributed to re-verify the changes. The most total dataset is output after several rounds of data labeling checks.

Continuously supervise the work of the annotator during the annotation process, solve problems arising in the annotation in a timely manner, and continuously adjust the annotation process and specifications according to the feedback in order to improve the objectivity and accuracy of the annotation.

2.2 Methods and Models

The preview classification model used in this paper is BERT , which is a pre-trained language representation model based on the Transformer model , and by pre-training on a large-scale unlabeled corpus , it learns a generalized language representation that can be applied to a variety of natural language processing tasks.Transformer's encoder understands each word's context through the mechanism of multiple heads of attention q, the outputs an embedding vector for each word.

The BERT model transformers framework allows for deep neural network based language extraction. It has strong characterization, context focus, and fine-tuning capabilities in news headline classification. It helps to better understand the semantics in news headlines and can provide efficient and accurate solutions for this task.

There are many reasons for choosing the BERT model over other traditional models because the BERT model has advantages that other models do not have. Compared to other models, BERT is bi-directional and it differs from traditional language models by its bi-directional context modeling capability. It takes into account contextual information from both the left and right directions, which better captures the relationships between words in a sentence and thus improves the understanding of context.In the pre-training and fine-tuning process, BERT adopts a two-stage training strategy, starting with unsupervised pre-training on a large-scale corpus, followed by supervised fine-tuning on specific tasks. This approach allows BERT to perform well on a variety of natural language processing tasks, while avoiding the large amount of time and resources consumed in training the model from scratch.BERT is also multi-task adaptable, BERT can be adapted to a variety of natural language processing tasks, such as text categorization, named entity recognition, question and answer, etc., with a single pre-training. This generality makes BERT a very powerful and flexible language representation model.Secondly BERT's contextual understanding is strong, BERT introduces both MLM(Masked Language Model) and NSP(Next Sentence Prediction) tasks during pre-training, which enables the model to better understand the semantics and structure in a sentence and improves the ability to understand information at the sentence level.BERT is also scalable in that the model can be augmented with additional layers or parameters to enhance its representational capabilities, making it highly scalable when dealing with text categorization tasks of different sizes.Finally its generality brings convenience to the research, as BERT is pre-trained on large-scale textual data, it has a certain degree of generality and can be applied to a wide range of text categorization tasks without the need to train from scratch.

2.3 Assessment of Indicators

For this study, two evaluation metrics were set up, and their metrics were analyzed and evaluated based on the running results, which can lead to the accuracy of the experiment. There are two evaluation metrics for the training set and the test set: training accuracy and training loss value, and test accuracy and test loss value, respectively.

In in this experiment, CrossEntropyLoss was used as the loss function of choice. The CrossEntropy loss function is commonly used for multi-categorization tasks, such as one of the commonly used loss functions when categorizing text.

Calculate the accuracy: In each batch, the output is first obtained by forward propagation through the model. Compare the output with the real labels and get the predicted labels by torch.argmax(). Calculate the number of predicted labels that are the same as the real labels and divide by the batch size to get the accuracy of the batch. Then sum up the accuracy of each batch to get the overall accuracy.

Calculate the loss value (loss): in each batch, calculate the cross-entropy loss between the model output and the real label, calculate the loss value and then sum up the loss value of each batch to get the overall loss value.
During the training process, losses are used to accumulate the loss values for each batch so that the overall loss can be calculated later. Accuracy is used to record the model's accuracy performance in the current epoch, which can be calculated at the end of each epoch based on the model's prediction results and real labels. In the training loop, optimizing the loss function to update the parameters of the model and monitoring the changes of the loss value and accuracy can help us understand the training of the model and make timely adjustments to improve the performance of the model.

3 Results

3.1 Analysis of Learning Rounds

Write the address of the written dataset to the model and start learning. Initially, the number of first learning rounds was set to five, and the learning rate was 5e-6, to see the training accuracy and loss values, and the testing accuracy and loss values.,the resultant data is shown in Table 1.

Table 1. First-time learning (learning rate 5e-6, 5 epochs).

Training Wheels	Training accuracy	Training loss	Test accuracy	Test loss
epochs1	0.289	1.381	0.294	1.393
epochs2	0.409	1.305	0.406	1.290
epochs3	0.615	1.140	0.630	1.140
epochs4	0.728	1.025	0.596	1.143
epochs5	0.777	0.971	0.711	1.030

Analyzing the data in Table 1, it can be seen that in five rounds of learning, with the gradual increase in the number of learning rounds, the training accuracy and test accuracy gradually increase and stay near 0.7; the loss value gradually decreases and stays around 1. It can be guessed that under the same learning rate, the more the number of learning rounds, the higher the accuracy rate and the loss value is about small.

The results show that the experimental results can be improved at a learning rate of 5e-6, and more rounds of learning will be performed below to verify the conjecture.

The second and third learning sessions are carried out below, with the number of learning rounds being 10 and the learning rate remaining constant at 5e-6, and the statistics are shown in Fig.1 below. From the line graph of the second learning (Fig. 1) , it can be seen that the accuracy rate tends to stabilize at 0.7 and the loss value tends to stabilize at 1 for the same learning rate. It can be speculated that the number of learning rounds has limited improvement on the accuracy rate , and the accuracy rate tends to stabilize after enough rounds of learning.

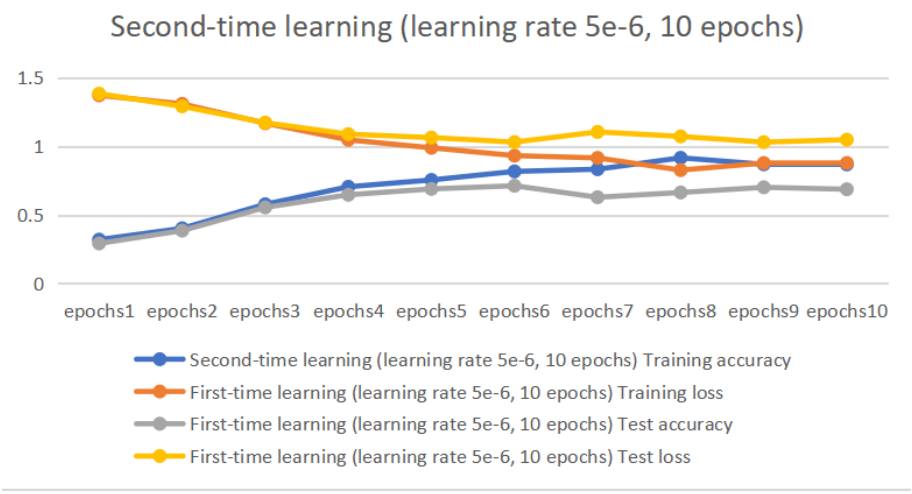

Fig. 1. Second-time learning (learning rate 5e-6, 10 epochs).

3.2 Adjustment of Dataset Labeling

The dataset labeling was improved with the same learning rate parameter and number of learning rounds and the resultant results were analyzed. The data was tallied and plotted as a line graph (e.g., Fig. 2). As can be seen in Fig.2, the third learning is comparable to the second, with the accuracy stabilizing at 0.7 and the loss value stabilizing at 1. It can be speculated that the number of learning rounds has limited improvement on the accuracy rate, and the accuracy rate tends to stabilize after enough rounds of learning. It shows that the dataset labeling has less impact on this experiment, and from the side, it also shows that the dataset labeling is more successful.

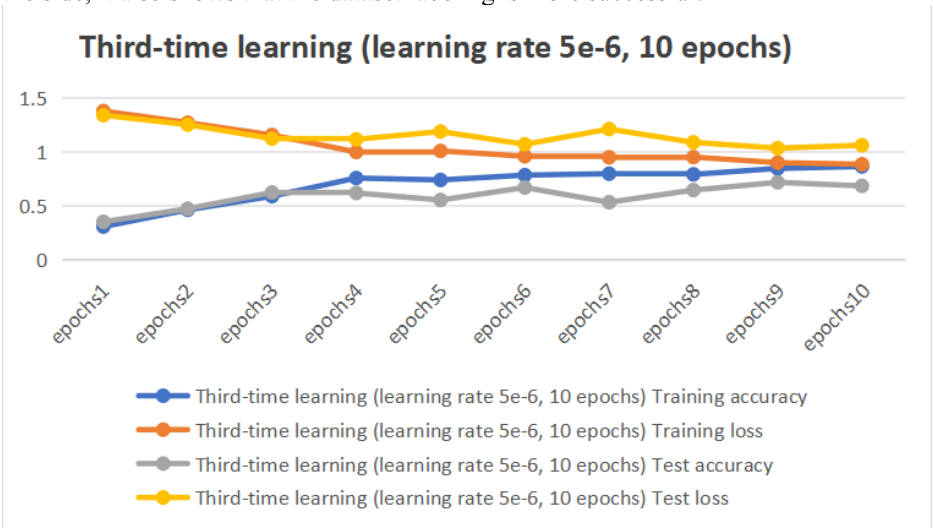

Fig. 2 Third-time learning (learning rate 5e-6, 10 epochs)

In the following, the research direction is shifted to the learning rate, and the number of learning rounds is kept constant to adjust the parameters to improve the learning rate.

3.3 Adjustment Parameters

The learning rate parameter was adjusted to 1e-5 and the number of learning rounds was 20 for testing. A line graph was drawn of the data statistics (e.g., Fig. 3).As can be seen from the line graph (Fig. 3), the accuracy of training after turning down the learning rate improves significantly to 0.8-0.9 with the increase in the number of learning rounds, but the test set accuracy decreases somewhat to 0.6-0.7.

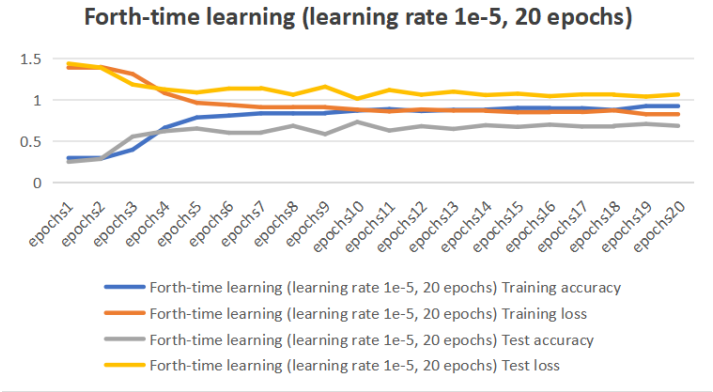

Fig. 3 Forth-time learning (learning rate 1e-5,20 epochs)

Continue to adjust the learning rate parameter to 1e-4 and the number of learning rounds to 10. The data statistics were plotted as a line graph (Fig. 4). From the results of the line graph (Fig. 4), it can be seen that after the learning rate parameter is tuned down again, the accuracy rate is directly reduced to 0.2 and the loss value rises to 1.4, which shows that it is not feasible to reduce the learning rate method.

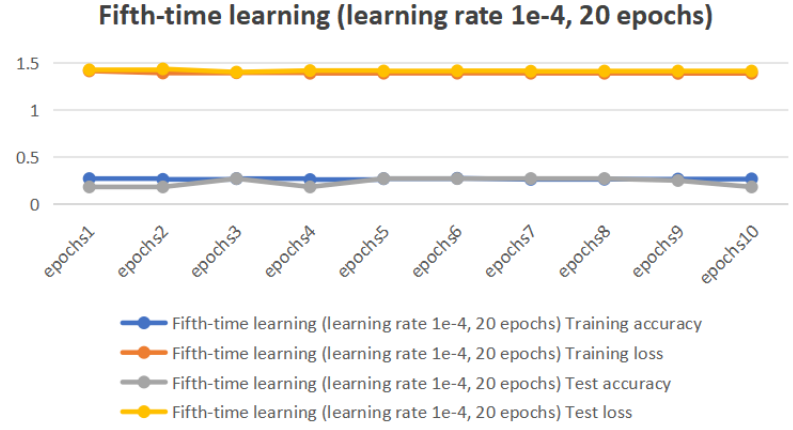

Fig. 4 Forth-time learning (learning rate 1e-5,20 epochs)

4 Discussion

4.1 Problem analysis

After adjusting the number of learning rounds and learning rate parameters separately to produce the results, the problem can be identified by analyzing and reviewing the information.

Learning rate is a very important hyperparameter in deep learning model training, which directly affects the convergence speed and final performance of the model. Too high or too low learning rate may lead to problems in the training process.

Too high a learning rate may lead to large fluctuations in the model parameters during the training process, or even lead to failure to converge and divergence.Too high a learning rate should also cause it to miss the optimal solution quickly, causing the model to skip areas near the optimal solution and thus fail to achieve optimal performance.

However, a low learning rate can lead to a very slow convergence of the model, requiring more iterations to achieve better performance.Moreover, a low learning rate is easy to fall into the local optimal solution, which may cause the model to hover at the local optimal solution and make it difficult to jump out of the local optimal solution to find the global optimal solution.

Therefore, when adjusting the learning rate, it is necessary to balance the convergence speed and performance. It is often recommended to gradually adjust the learning rate until the optimal value is found by experimenting and observing the performance on the validation set. Meanwhile, common learning rate decay strategies also help to dynamically adjust the learning rate during the training process to balance the convergence speed and performance performance of the model.

The reasons for the problem of high correct rates in training and low correct rates in testing that occurred in the fourth study will be analyzed below.

The situation of high training correctness and low test correctness is often referred to as "overfitting", which is a problem that deep learning models often face. Overfitting refers to the phenomenon where a model performs well on a training set but performs poorly on unseen data such as a test set. Some of the possible reasons for high training correctness and low test correctness are numerous.

Insufficient data or uneven data distribution can lead to unstable correctness. Inconsistent data distribution between the training and test sets, or too few samples in the training set, results in models that cannot be generalized to the test set.Excessive model complexity can also contribute to the problem. Models with too many parameters or layers make it easy to memorize the noise in the training set instead of the true pattern.And setting the learning rate too high may also cause the model to skip the optimal solution during training and fail to reach convergence.A training process that is too long or over-training the model can potentially cause the model to overfit the training set.Poor feature selection can also affect the correctness rate, with features that are too complex or contain noise, making it difficult to generalize the model to new data.

4.2 Problems Solved

The following are possible solutions to the various problems that have arisen in this paper.

For the number of learning rounds, it should be set appropriately, between 10-20 is more appropriate. For the learning rate parameter adjustment, the problem of insufficient data or uneven data distribution can be improved by increasing the amount of training data, improving the quality of datasets, and performing data enhancement to improve the generalization ability of the model.For the model complexity is too high, you can simplify the model structure, increase the regularization term, use Dropout and other methods to prevent overfitting. For the learning rate is too high, the learning rate can be appropriately reduced and different learning rate scheduling strategies such as learning rate decay can be tried.If the training process is too long, an early stop strategy can be used to stop training when the performance on the validation set is no longer improving to avoid overfitting. If the features are not selected properly, feature selection, dimensionality reduction or feature engineering should be performed to ensure the quality and relevance of the input features.

5 Conclusion

In this study, the application of Bert model on the task of Chinese news headline classification is thoroughly discussed and analyzed. The differences between the Bert model and traditional methods in text classification are compared, reflecting the advantages and performance of Bert in Chinese news headline text classification. Meanwhile, the performance differences between the Bert model under different learning rate parameters, number of learning rounds, and different dataset labeling accuracy settings are analyzed, as well as its advantages and difficulties in dealing with Chinese text data are discussed.

The Bert model achieves better performance on the news headline classification task. Compared with traditional methods based on sequence annotation or bag-of-words models, Bert achieves higher accuracy and better generalization ability on the news headline classification task. The performance varies under different parameter settings, and the results show that the appropriate parameter settings have a significant impact on the performance of the Bert model, which provides some suggestions and guidance for the subsequent users.The Bert model has certain advantages when dealing with Chinese text data, and it achieves good results in the news headline classification task.

Varying the number of learning rounds can show after how many rounds the learning accuracy stabilizes, with all other variables held constant. Changing the learning rate parameter can greatly affect the learning correctness and loss values. Adjusting the dataset labeling with the resultant correctness being essentially unchanged can also yield different learning results.

Many shortcomings were found in this experiment, which should be addressed in future improvements as mentioned above. Improve the generalization ability of the model; improve the model complexity to prevent overfitting. Adjust the learning rate

and try different learning rate scheduling strategies; adjust the training process to avoid overfitting; improve the dataset labeling accuracy to improve the learning accuracy.

From the results of this study, it can be introduced that the model can improve the efficiency and mode of receiving news, can classify and organize a large amount of news in a timely manner and archive it quickly, and can analyze the public opinion orientation and social hotspots according to the categories. It can realize the personalized classification of news and push the received news by category. It reveals the potential and advantages of the Bert model in the task of news headline categorization, and provides useful insights and guidance for future research and practice in the field of natural language processing, which is of great theoretical and applied significance.

References

1. Li J.W.: Research and Implementation of News Text Classification System Based on Deep Learning. Beijing University of Posts and Telecommunications,(2019).
2. Miao J.:Zhang Y.S.,Li J.L.:Classification of Chinese News Headlines Based on Bidirectional Encoder Representations from Transformers(BERT)[J]. Computer Engineering and Design, **43**(08):2311-2316(2022).
3. Xu P.: Research on news text classification method based on deep learning,Nanjing University of Information Engineering, (2023).
4. Wu Y.: Research and Application of Deep Learning Based Text Classification of News Headlines, Yangzhou University, (2024).
5. Lu Z.B.: Research on Text Classification Algorithm for Chinese News Headlines Based on Deep Learning , Southwest University, (2023).
6. Qian A.B., Jiang L.: Automatic categorization of Chinese news web pages based on headlines, Modern Library Intelligence Technology, (10): 59-68(2008).
7. Jing W., Bailong Y. News text classification and recommendation technology based on wide & deep-bert model, 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE). IEEE, 209-216 (2021).
8. Tan Z., Chen B., Fang W.:Analysis and Application of Financial News Text in Chinese Based on Bert Model,Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference, 35-39 (2020).
9. Ambalavanan A K, Devarakonda M V.: Using the contextual language model Bidirectional Encoder Representations from Transformers(BERT) for multi-criteria classification of scientific articles, Journal of biomedical informatics,**112**, 103578 (2020).
10. Rai N, Kumar D, Kaushik N, et al. Fake News Classification using transformer based enhanced Long Short-Term Memory(LSTM) and Bidirectional Encoder Representations from Transformers(BERT), International Journal of Cognitive Computing in Engineering, **3**: 98-105 (2022).
11. Zhang Q, Li J, Jia Q, et al.: UNBERT: User-News Matching Bidirectional Encoder Representations from Transformers(BERT) for News Recommendation,IJCAI., **21**, 3356-3362(2021).
12. Lin D.P.: Research on News Text Classification Based on Deep Learning, Beijing Institute of Printing, (2024).
13. An J.X.,Jiang S.C.: A review of word vector modeling research for natural language processing. Computer Technology and Development, **33** (12): 17-22 (2023).