Using Bert+TextCNN Model to Realize News Classification and Recognition

Zhang Yang 36920221153150 [1] **Fu Honghao 36920221153078** [1] **Pan Jiawei 31520221154219** [1] **He Zhifeng 36520221151953** [1]

[1]Xiamen University
36920221153150@stu.xmu.edu.cn, 36920221153078@stu.xmu.edu.cn, 31520221154219@stu.xmu.edu.cn, 36520221151953@stu.xmu.edu.cn

Abstract

With the development of modern information technology such as big data and cloud computing, traditional paper documents are rapidly changing to electronic and digital, and the basic methods of document management such as classification and retrieval are also changing. How to effectively manage these information and obtain valuable content from it is a major challenge in text processing. Text classification is the basic work to solve this challenge. Classifying a given text through an algorithm is an important basis for subsequent text automatic operation and processing. Therefore, it is particularly important to find a fast and high-precision text classification algorithm. In this paper, Naive Bayesian Model and Bert+TextCNN model can be used to divide specific categories into nine categories: finance, real estate, education, science and technology, military, automobile, sports, games and entertainment, so that readers can quickly find the types of news they want to read.

Introduction

News refers to the name of information spread through newspapers, radio, television and other media channels. It is a style of writing that records society, spreads information and reflects the times. The concept of news can be divided into broad sense and narrow sense. In its broad sense, in addition to the comments and special articles published in newspapers, radio, the Internet, and television, the commonly used texts belong to news, including news, communications, features, sketches, and so on. The narrow sense of news refers specifically to news, which is to quickly and timely report the recent and valuable facts at home and abroad in a concise and concise way to let others know. With the vigorous development of the Internet, the number of all kinds of news also shows explosive growth. If readers only rely on manual search, it will be difficult to find the news they are interested in. Therefore, using an effective and fast model can not only facilitate different reading groups to quickly select the news they are interested in according to their needs, but also effectively meet the scientific retrieval needs of massive news materials.This model is divided into two stages, using different classification models for different classification tasks. The task of the first stage is to use Naive Bayesian Model to divide all data into classes with specific categories and other classes. In the second stage, Bert+TextCNN model is used to divide the specific categories into nine categories: finance, real estate, education, science and technology, military, automobile, sports, games and entertainment.

Classification is a basic problem in the field of data analysis and machine learning. Text classification has been widely used in many aspects such as network information filtering, information retrieval and information recommendation. Data driven classifier learning has been a hot topic in recent years, with many methods, such as neural networks, decision trees, support vector machines, naive Bayes, etc. Compared with other more complex classification algorithms designed carefully, Naive Bayes classification algorithm is one of the classifiers with better learning efficiency and classification effect. Intuitive text classification algorithm, also the simplest Bayesian classifier, has good interpretability. Naive Bayesian algorithm is characterized by the assumption that all features are independent of each other, and each feature is equally important. But in fact, this assumption does not hold true in the real world: first, the inevitable connection between two adjacent words cannot be independent; Secondly, for an article, some of the representative words will determine its theme. It is not necessary to read the whole article and view all the words. Therefore, it is necessary to adopt appropriate methods for feature selection, so that naive Bayesian classifier can achieve higher classification efficiency.

Naive Bayesian Model is a classification method based on Bayesian theorem and independent hypothesis of feature conditions, and is one of the most widely used classification algorithms. Naive Bayesian method is simplified based on Bayesian algorithm, that is, the attributes are assumed to be conditionally independent when the target value is given. That is to say, no attribute variable has a large proportion of the decision-making results, and no attribute variable has a small proportion of the decision-making results. Although this simplified method reduces the classification effect of Bayesian classification algorithm to a certain extent, it greatly simplifies the complexity of Bayesian method in practical application.

Bert (Bidirectional Encoder Representation from Transformers) is a pre trained language representation model.

It emphasizes that the traditional one-way language model or shallow splicing of two one-way language models is no longer used for pre training, but the new MLM (masked language model) is used to generate in-depth two-way language representation.

Bert showed amazing results in the top level test of machine reading comprehension, SQuAD1.1: he surpassed humans in all aspects of the two indicators, and achieved SOTA performance in 11 different NLP(nature language processing) tests, including pushing the GLUE benchmark to 80.4% (7.6% absolute improvement), and the accuracy of MultiNLI reached 86.7% (5.6% absolute improvement), becoming a milestone model achievement in the history of NLP development and the most outstanding NLP language representation model at present. Therefore, this project uses Bert model to represent news text in vector.

Related work

Text Classification

Classification is a fundamental problem in the field of data analysis and machine learning. For a given text, the model outputs the category information of the text. Text classification has been widely used in many fields such as Web information filtering, information retrieval and information recommendation. Data-driven classifier learning has been a hot topic in recent years. There are many methods, such as neural networks, decision trees, support vector machines, naive Bayes, etc.

Classification Model

Compared with other carefully designed more complex classification algorithms, Naive Bayes classification algorithm is one of the classifiers with better learning efficiency and classification effect. The intuitive text classification algorithm is also the simplest Bayesian classifier, which has good interpretability. The characteristics of the naive Bayes algorithm are that the appearance of all features is independent of each other, and each feature is equally important. But in fact, this assumption does not hold true in the real world: first, the necessary relationship between two adjacent words cannot be independent; Secondly, for an article, a few representative words can determine its topic, it is not necessary to read the whole article and look at all words. Therefore, it is necessary to use an appropriate method for feature selection, so that the naive Bayes classifier can achieve higher classification efficiency.

The TextCNN model was first proposed by Yoon Kim to use convolutional neural networks to deal with NLP problems. Compared with the traditional RNN, LSTM and other models in NLP, CNN can extract important features more efficiently, and these features play an important role in classification. TextCNN is a very classic model in NLP. Through verification experiments and the industry consensus, in text classification tasks, CNN model has been able to get good results. Although the effect may be slightly worse than RNN on some data sets, CNN model training is more efficient. Therefore, it is generally believed that CNN model is an ideal model with both efficiency and quality in text classification tasks.

Pre-trained Language Model

Pre-trained language model has shown good results in extracting text features. Through pre-training on large-scale corpus, the semantic information of the text is learned, and the pre-trained model is obtained. Then it is applied to downstream tasks to provide richer information for downstream tasks. Bert (Bidirectional Encoder Representation from Transformers) is a pre-trained language model. Instead of using the traditional one-way language model or the shallow concatenation method of two one-way language models for pre-training, it uses a new MLM (masked language model), which can generate deep bidirectional language representation.

Bert performs surprisingly well on the SQuAD1.1 Top level machine Reading Comprehension test: We outperform humans on all two metrics and achieve SOTA performance on 11 different NLP benchmarks, including driving the GLUE benchmark to 80.4

Proposed Solution

Our proposed model is divided into two stages in total, and different classification models are used for different classification tasks. For the first stage: use the Naive Bayes model to classify all the data into specific classes (finance, real estate, education, technology, military, automotive, sports, games, entertainment) and other classes. For the second stage: Bert is used to extract vector representations of the text, and then TextCNN is used to divide the classes with specific categories into nine categories: finance, real estate, education, technology, military, automobile, sports, games, and entertainment.

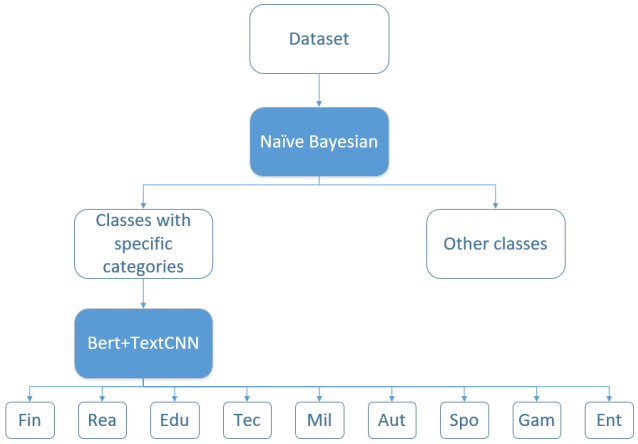

Figure 1: Our proposed model flowchart. 'Fin', 'Rea', 'Edu', 'Tec', 'Mil', 'Aut', 'Spo', 'Gam' and 'Ent' represents finance, real estate, education, technology, military, automotive, sports, games and entertainment respectively.

Naive Bayesian

We adopt the naive Bayes model for classification in the first stage . The classification task required in this project is similar to the spam classification task, and spam and "other class" characteristics cannot be fully listed. Naive Bayes model has achieved good results in spam classification, so the naive Bayes model is used to complete the classification of "other classes". The principle of naive Bayes model is as follows: For a given item to be classified, calculate the occurrence probability of each class under the condition of the occurrence of this item, and consider that the item belongs to the class with the largest probability.

Naive Bayesian Model is a classification method based on Bayes theorem and conditional independence assumption of features, which is one of the most widely used classification algorithms. In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on Bayes' theorem under the assumption of strong (naive) independence between features. The naive Bayes method is simplified from the Bayesian algorithm, that is, the attributes are assumed to be conditional independent of each other when the target value is given. That is to say, no attribute variable occupies a large proportion for the decision results, and no attribute variable occupies a small proportion for the decision results. Although this simplification method reduces the classification effect of the Bayesian classification algorithm to a certain extent, it greatly simplifies the complexity of the Bayesian method in the actual application scenario.

Naive Bayes classification is a method based on Bayes theorem and assuming that the feature conditions are independent of each other. It first learns the joint probability distribution from input to output through the given training set and assumes that the feature words are independent, and then based on the learned model, it input X to find the output Y that maximizes the posterior probability.

Let the sample data set$D = \{d_1, d_2, \cdots, d_n\}$, for the sample data characteristic attribute set $X = \{x_1, x_2, \cdots, x_d\}$, the class variable $Y = \{y_1, y_2, \cdots, y_m\}$, that is D can be divided into y_m categories. And $x_1, x_2, \cdots, x_d$ is independent of each other and random, then the posterior probability $P_{post} = P(Y|X)$ of y can be calculated from Equation 1.

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{1}$$

Naive Bayes is based on the independence of each feature, and Equation 1 can be further expressed as Equation 2 when the given class is y.

$$P(X|Y = y) = \prod_{i=1}^{d} P(x_i|Y = y) \tag{2}$$

From Equations 1 and 2, the posterior probability can be calculated as Equation 3

$$P_{post}(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(x_i|Y)}{P(X)} \tag{3}$$

Since the size of $P(X)$ is fixed, we can only compare the numerator part of Equation 3 when comparing posterior probabilities. Therefore, the naive Bayes calculation Equation 4 for a sample data belonging to the class y_i can be obtained

$$P(y_i|x_1, x_2, \cdots, x_d) = \frac{P(y_i)\prod_{i=1}^{d} P(x_j|y_i)}{\prod_{j=1}^{d} P(x_j)} \tag{4}$$

Bert

BERT is a multi-layer bidirectional Transformer encoder based on fine-tuning, where the Transformer is the same as the original Transformer. We use Bert model for vector representation of news text, and learn feature representation for text by running self-supervised learning method on the basis of massive corpus. The BERT pre-trained model is divided into the following three steps:

- Given the representation input to Bert, the model can explicitly represent a sentence or sentence pair (e.g., question and answer) for different tasks. For each token, its representation is generated by adding its corresponding token embedding, segment embedding and position embedding, as shown in Figure 2.
- In the Bert model, the Transformer is trained in the way of Mask-LM, and part of the words of the input sentence are randomly removed as the label to be predicted, and then the bidirectional deep Transformer model is trained. Therefore, in order to be consistent with the subsequent task, we need to input the original word or some random word in a certain proportion of the word position to be predicted. Of course, Bert is less efficient than a normal language model because only a subset of words are used for training. The authors also point out that it takes more training steps for Bert to converge.
- In addition, the Bert model adds an additional sentence-level continuity Prediction task Next Sentence Prediction on the basis of the bidirectional language model to predict whether the two ends of the text input to BERT are continuous text. This task allows the model to better learn the relationship between consecutive pieces of text. Specifically, some sentence pairs A and B are selected, where 50

Figure 3 shows the Bert model applied to downstream tasks. We can easily use Bert to extract text representations for subsequent classification tasks.

TextCNN

Figure 4 shows the overall structure of TextCNN. The input data is first passed through an embedding layer to obtain the embedding representation of the input sentence, then through a convolution layer to extract the features of the sentence, and finally through a fully connected layer to obtain the final output.

- Embedding layer, the main role is to encode the input natural language into a distributed representation. Pretrained word vectors can be used or a set of word vectors can be trained directly in the process of training TextCNN.

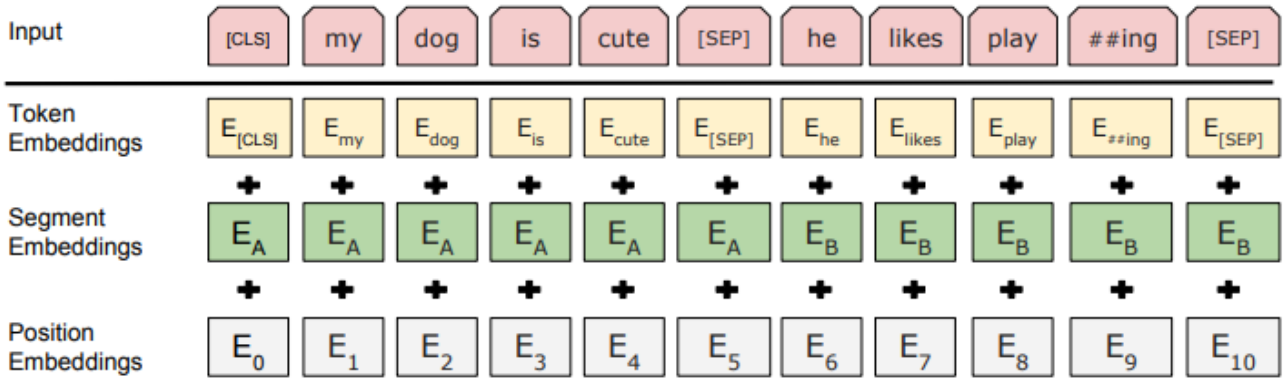

Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings

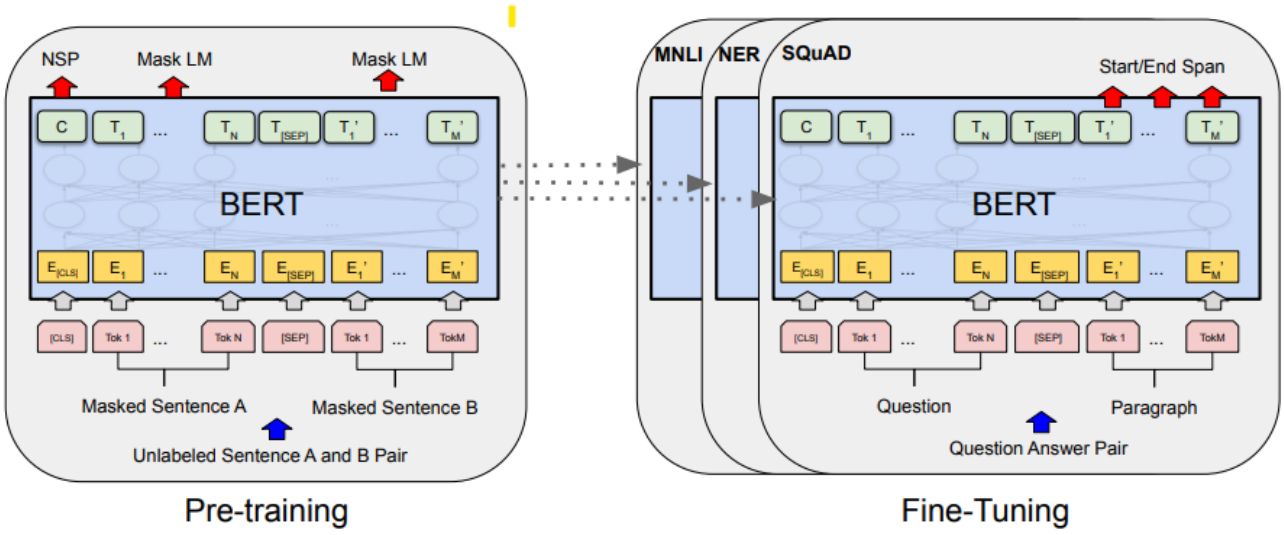

Figure 3: Overall pre-training and fine-tuning procedures for BERT.

- This layer is mainly through convolution, extracting different n-gram features. The input sentence or text, after passing through the embedding layer, will be transformed into a two-dimensional matrix. The next convolutional work is done on a two-dimensional matrix. The size of the convolution kernel is usually set to $n \times |d|$, where n is the length of the convolution kernel and d is the width of the convolution kernel, which is the same as the dimension of the word vector, that is, the convolution is only performed along the text sequence. There can be multiple choices, such as 2, 3, 4, 5, and so on. In TextCNN network, multiple kernels of different types need to be used at the same time, and at the same time, there can be multiple kernels of each size.

- The Max pooling layer takes the maximum value of several one-dimensional vectors obtained after convolution, and then concatenates them together as the output value of this layer.

- The fully connected layer, concatenates a layer after the max-pooling layer, as the output result. To improve the learning ability of the network, multiple fully connected layers can also be concatenated.

Experiments

We use crawler technology to crawl data of corresponding categories from news websites such as Baidu News, Sina News, NetEase News, Tencent News, Renmin website, Xinhuanet and China National Broadcasting Network, and clean the crawled data and add it to the training data set to eliminate the imbalance of data categories. The balanced data distribution is shown in figure 5.

In the experiment, the naive Bayes model was used to identify other types of news. The training data set was a total of 80,000 news data sets, of which 20,000 were ”other classes” and about 60,000 were non-other classes. After that, the Bert+TextCNN model was used to identify specific types

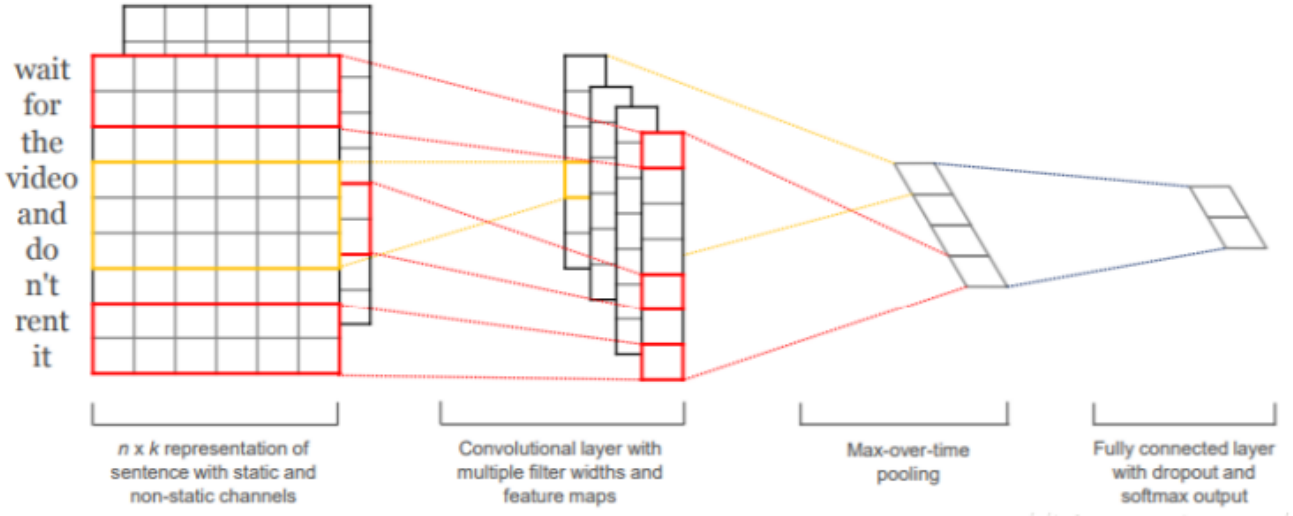

Figure 4: The input data is first passed through an embedding layer to obtain the embedding representation of the input sentence, then through a convolution layer to extract the features of the sentence, and finally through a fully connected layer to obtain the final output.

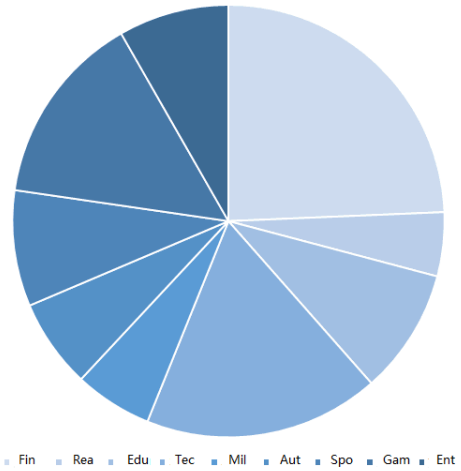

Figure 5: News text data distribution. 'Fin', 'Rea', 'Edu', 'Tec', 'Mil', 'Aut', 'Spo', 'Gam' and 'Ent' represents finance, real estate, education, technology, military, automotive, sports, games and entertainment respectively.

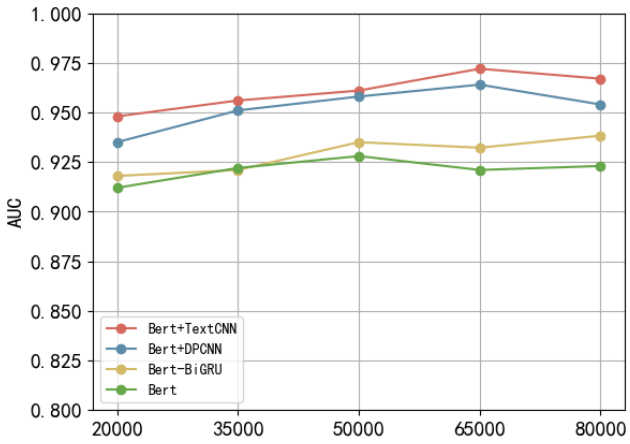

Figure 6: Our experimental results in different dataset lengths, Bert+TextCNN is our proposed method and compared with several other methods.

of news. The training data set recorded a total of 60,000 data, and each type of data was divided into training set, validation set and test set according to 8:1:1. Thirty words, including the title, are fed into the model for training. As a comparison, we select three models with better performance, Bert, Bert+DPCNN and Bert+BiGru. The experimental results are shown in Figure 6.

Experiments show that our proposed method significantly outperforms the Bert model and the Bert combined with BiGru model, and slightly outperforms the Bert combined with DPCNN model.

Conclusion

According to the characteristics of other types of news, this paper proposes the concept of category typical word and the method of finding this word, and uses the category typical word to improve the classification effect of the model. Our model consistently outperforms the other methods compared across different training dataset lengths. In the future, we will consider adding sentiment analysis to the text and make a more fine-grained analysis of the news text.

References

Chen, Y. 2015. *Convolutional neural network for sentence classification*. Master's thesis, University of Waterloo.

Deng, L.; Wang, J.; Liang, H.; Chen, H.; Xie, Z.; Zhuang, B.; Wang, S.; and Xiao, J. 2020. An iterative polishing framework based on quality aware masked language model for Chinese poetry generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7643–7650.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Han, C.; Lin, Q.; Man, Z.; Cao, Y.; and Wang, H. 2021. Automated Classification of Textual SPECT Diagnostic Reports with TextCNN Model. In *Journal of Physics: Conference Series*, volume 1792, 012023. IOP Publishing.

He, W. 2018. *Text Classification Algorithm Research Based on Naive Bayes*. Master's thesis, Nanjing University of Posts and Telecommunications.

Liang, T.; Yang, X.; Wang, L.; Zhang, Y.; Zhu, Y.; and Xu, C. 2017. Review on Research and Development of Memory Neural Networks. *Journal of Software*, 28(11): 2905–2924.

Liu, H.; Perl, Y.; and Geller, J. 2020. Concept placement using BERT trained by transforming and summarizing biomedical ontology structure. *Journal of Biomedical Informatics*, 112: 103607.

Mustofa, R.; and Prasetiyo, B. 2021. Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on# newnormal hashtag in twitter. In *Journal of Physics: Conference Series*, volume 1918, 042155. IOP Publishing.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhaoye, X.; Xiaoqun, L.; and Peijie, S. 2021. Hybrid Chinese text classification model based on pretraining model. In *Journal of Physics: Conference Series*, volume 1961, 012002. IOP Publishing.

Zhou, F.; and Li, R. 2018. Convolutional Neural Network Model for Text Classification Based on BGRU Pooling. *Computer Science*, 45(06): 235–240.