Research on Text Classification Model Based on Self-Attention Mechanism and Multi-Neural Network

Xiaolin Wang[1], Yue Chen[2], Wei Liu[2], Weipeng Tai[1,*]

[1]*Engineering Research Institute in Anhui University of Technology, China*
[2]*School of Computer Science and Technology in Anhui University of Technology, China*

Abstract

The traditional text classification algorithms based on deep learning include RNN and CNN and their variants. The too long sequence of RNN is easy to produce gradient explosion and the gradient disappears, so the long-distance dependence of the text cannot be extracted; CNN focuses on the local features of the sentence rather than the global structure of the sentence. In response to this problem, this paper proposes a text classification model (Self-attention and Multiple Neural Networks Unit based Text Classification SMNN) that integrates the self-attention mechanism and multiple neural networks. This structure uses a word embedding model based on the self-attention mechanism that focus on the important parts of the text to generate a global representation of the text, uses CNN to extract the local semantic features of the text at multiple granularities through different convolution kernel sizes and k-max-pooling, and uses BiLSTM with skip connections and pooling layers to extract the long-distance dependencies of the text for obtaining the global representation of the text. Then it fuses the global features and local features. Lastly, it classifies the text information through the softmax classifier. The experimental results on the text data set show that the SMNN model has better text classification accuracy which proves that the SMNN model has obvious advantage and value compared with the traditional text classification model.

Keywords

SMNN; BiLSTM; CNN; text; classification; semantic features

1. Introduction

In the early text classification algorithms, statistical-based learning models dominated[1], such as Support Vector Machine (SVM), Naive Bayes and k-nearest neighbor algorithms. These text classification algorithms have obvious shortcomings, the context information and sentence structure of the sentence are not considered, and the relationship between words is ignored, resulting in low learning ability and poor generalization ability of the model, and it is difficult to make accurate predictions. At the same time, statistical-based learning methods are used in feature engineering. It consumes a lot of time, so the concept of deep learning is proposed.

Hinton[2] proposed a new concept in 2006 to solve the problems of machine learning algorithms. They proposed the concept of deep learning to solve the problems of poor representation and generalization ability of previous machine learning algorithms. Since then Later, deep learning continued to develop with its powerful feature selection and information extraction capabilities.Text classification algorithms based on deep learning gradually began to replace traditional machine learning methods, such as convolutional neural networks and recurrent neural networks, and their models were improved, and then to the transformer model based on the attention mechanism proposed by Google, deep learning has made remarkable achievements in the field of text classification algorithms.

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China
wxl@ahut.edu.com (Xiaolin WANG), 2426922665@qq.com (Yue CHEN), 453389711@qq.com (Wei LIU), taiweipeng@ahut.edu.cn* (Weipeng TAI)

In this paper, a word embedding model based on self-attention mechanism is used to obtain the global representation of text, and CNN is used to extract local semantic features of text at multiple granularities through different convolution kernel sizes and k-max-pooling, and BiLSTM with skip connections is used. And the pooling layer to obtain the global semantic features of the text, and then fuse the local features and global features, and classify the text information through the softmax classifier. This structure can capture the local and global features of the sentence at the same time, and improve the effect of text classification.

2. Related work

The text classification method of deep learning has been developed to a great extent since 2013, and gradually replaced the existing traditional machine learning methods. Many researchers have participated in the research and proposed a large number of deep learning models for text classification. Compared with traditional machine learning models, deep learning-based models can effectively solve the high-dimensionality and matrix sparse problems of text feature vectors. Now the main research deep learning models can be divided into: RNN-based text classification models, CNN-based classification models, transformer-based, attention-based text classification models, and PLM-based text classification models.

2.1.CNN network

The neural network based on CNN has achieved great success in the field of images, so some researchers and scholars in colleges and universities proposed to apply the CNN network to the related tasks of text classification. The earliest CNN network model used for text classification tasks is Kalchbrenner[3] proposed the DCNN network model, which mainly captures sentences by convolution filling as the way of same, that is the alternating structure of wide convolution and k-max dynamic pooling. Feature map, which can well capture the local and global features of the sentence. Compared with DCNN, the textCNN proposed by Kim[4] is a simpler network model. It uses single-layer convolution and maximum pooling to extract and represent sentence features. It can extract the information of each word in the text. complete feature. For the improvement of DCNN, Johnson[5] proposed a DPCNN model that improves the performance of the model by increasing the network depth, which improves the performance of the model to a certain extent. Conneau[6] also made further improvements to DCNN and proposed the VDCNN network model. Its innovation lies in applying convolution and pooling with small steps to character-level vectors. The classification performance will increase, but with the increase of the network, the time complexity of model training increases and the defect of semantic loss caused by the deepening of the network is particularly obvious.

2.2.RNN network

The RNN text classification model is a deep learning network structure proposed by Jordan[7], which can better learn text features, understand text meaning, and capture the global features of text sentences. The RNN neural network model mainly learns text sentence sequences through recursive calculation. feature. However, there are its own defects in the process of RNN learning. The obvious problem is that there are gradient disappearance and gradient explosion in the process of recursive calculation. The above-mentioned problems Zhang[8] and others refer to the long short-term memory network model (long short-term memory, LSTM) improved on the basis of the RNN network structure based on the sentence state in the text classification task and the performance has been greatly improved, which effectively alleviates the problem above problems. Since the structure of the sentence has contextual information, but LSTM only considers the one-way information of the sentence (only the above information of the sentence is considered but not the following information of the sentence), the deviation of LSTM in semantic understanding will affect the accuracy of the model.

2.3.Self-attention network and others

The attention mechanism has made leap-forward achievements and breakthroughs in the field of deep learning in recent years. Whether it is in the field of natural language or images, many scholars and researchers are committed to combining attention mechanism and neural network for text classification. Research. The core of the attention mechanism is to obtain and learn the most useful information under limited resources. The essence of the attention mechanism queries the mapping of a series of key-value pairs, and calculates the similarity between the keys and values in the key-value pairs to obtain the weight, which improves the accuracy of text classification by assigning different weights to the text content. The first application of the attention mechanism was used in machine translation by Bahdanau[9] and achieved very good results.

Since then, Luong[10] Defined global attention and local attention in machine translation. Vaswani[11] proposed a sentence representation method based on Self-Attention mechanism for machine translation, which greatly improved the effect of machine translation. Due to the excellent performance of the self-attention mechanism in the field of translation, many scholars have considered combining the self-attention mechanism with deep learning models.For example, Jia Hongyu[12] proposed a text classification model RCNN_A that combines self-attention mechanism and recurrent neural network; Xinqiang[13] proposed to combine self-attention mechanism and BiLSTM to capture local features and global features of sentences features, so that the classification effect of the network model has been improved. The self-attention mechanism has its advantages and disadvantages. The advantage is that it can capture the relevance of text without other information. The disadvantage is that it cannot capture the timing information of sentences. Therefore, it is necessary to add positional encoding to the self-attention mechanism to better capture the features of sentences.

Early pre-trained language models are usually only used as text word embeddings and will not be used for text classification tasks, such as the commonly used pre-trained language models Word2vec and Glove[14][15]. The role of pre-trained language model models on text classification tasks begins with the transformer model proposed by Google. Many pre-trained language models based on transformer models not only focus on the word embedding of text, but also act on downstream tasks. Such as the OpenGPT model proposed by Radford[16] and the Bert model proposed by Devlin[17]. Bert has achieved excellent results in multiple natural language processing tasks such as e-commerce[18], medical[19] and finance by learning contextual representations through a bidirectional network structure. Subsequently, many researchers further improved the Bert model, such as the BERT-wwm model[20], the ALBERT model[21], the Bert-CNN[22] and the BiLSTM-CRF[23] model proposed by Shi Zhenjie, Dong Zhaowei.

3. SMNN network

The SMNN model proposed in this paper obtains the representation of global and local semantic feature vectors through the fusion of Word Embedding and self-attention mechanism, and finally obtains the results of text classification through a multi-model fusion mechanism. The network structure of the SMNN model is shown in the figure 1, which consists of an input layer, an embedding layer, a CNN layer, a BiLstm layer and a fusion output layer.

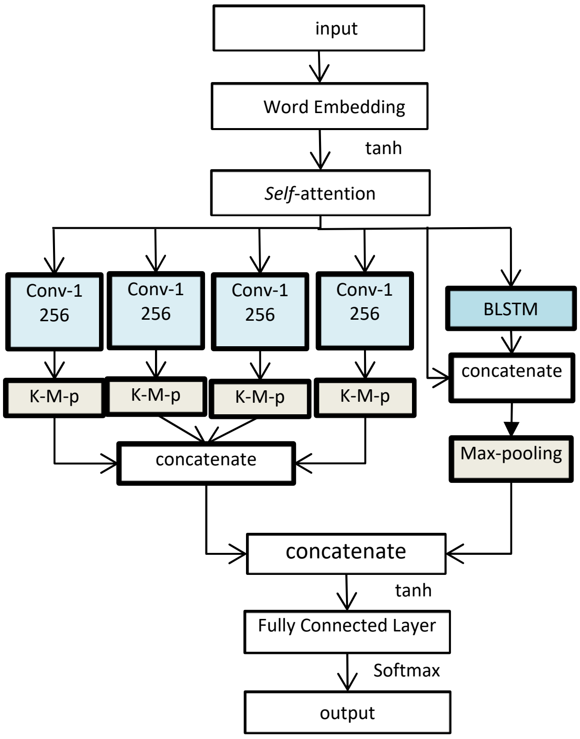

Figure 1. SMNN network

3.1. Embedding layer

Word embedding is a standard pre-trained language model that generates word vectors. In the input layer, each word in the sentence is represented by a one-hot vector, and then the data of the input layer is input into the word embedding to generate the distributed word vector of the sentence to realize the matching Dynamic encoding of words. The traditional word vector representation model cannot take into account the degree of mutual attention between the current word of the sentence and the words in other positions, so as to affect the representation of the word vector. Different from the traditional text classification model, the SMNN text classification model uses the word embedding mechanism based on the self-attention mechanism to obtain the feature representation vector of each word in the sentence. First, the word embedding is used to generate the distributed word vector, and then the activation function Tanh is used. The processing is to speed up the calculation and training of the model, and then use the self-attention mechanism to calculate the degree of association between each word and other words in the text, and finally output the representation M of the word vector. For example, a sentence vector containing contextual information is generated in the embedding layer and processed by the activation function tanh to obtain H, and h_t^* represents the word vector corresponding to the t-th word of H. Many staff, authors and scholars consider that each word pair is for the degree of influence of global semantics is different, so it is considered to introduce a self-attention mechanism to assign different weights to each sentence, and use the weights to determine the influence of the current word on the semantics of the text and reflect the importance of the word to the sentence. and way to obtain the representation M of the global feature vector of the text:

$$u_t=tanh(W_w h_t^* + b_w) \tag{1}$$

$$a_t=\frac{exp(u_t^T u_w)}{\sum_{t=0}^{n} exp(u_t^T u_w)} \tag{2}$$

$$M=\sum_{t=0}^{n} a_t h_t^* \tag{3}$$

Where $W_w$represents the parameters of model training,$b_w$represents the bias term, u_t^Trepresents the transpose of u_t, $u_w$is a randomly initialized context vector of a model, and a_t represents weight of words in the t-th moment of the input sequence after normalization.

3.2. CNN layer

Convolutional neural network (CNN) is a good local feature extractor, which captures the local features of data by controlling the size of the convolution kernel. Therefore, Kim et al. proposed the textCNN network structure to use the word vector generated by the embedding layer as input. As shown in Figure 2, compared with the traditional textCNN model, the SMNN model takes the feature matrix M $\in R^{h\times d}$ generated by the attention mechanism model training as input, and then uses a one-dimensional window size of 1, 2, 3, and 4. The convolution kernel performs vertical convolution on M to obtain feature maps of different granularities of text. In the SMNN model, the Selu activation function is used to process the feature maps to improve the classification effect and prevent the death of neurons caused by relu processing. At the same time, in order to prevent max- Pooling feature selection results in feature loss. The feature maps generated by different convolution kernel volumes in SMNN use K-Max-Pooling maximum pooling operation to select k features (k=2, indicating that the two most important features are selected), thus generating The feature representation contains rich local feature information of text, and the final output vector $\boldsymbol{F_l}$, $\boldsymbol{F_l}$ is the local feature representation of the CNN layer. The expression of the feature map is C=[c_1,c_2,⋯,c_{n-h+1}], and the calculation process of each element c_i in the feature map is as follows:

$$c_i=f(w_c * M_{i:i+h-1}+b_c) \tag{4}$$

Among them, f represents the activation function (*selu*), $w_c \in R^{h\times d}$ is the convolution kernel, * represents the product operation of the elements in the matrix, $b_c$represents the bias term, and $h\in\{1, 2, 3, 4\}$ represents the convolution kernel size , $M_{i:i+h-1}$ means to take the data from row i to row *i+h-1* of *M*, where the value range of i is [1,*n-h+1*].

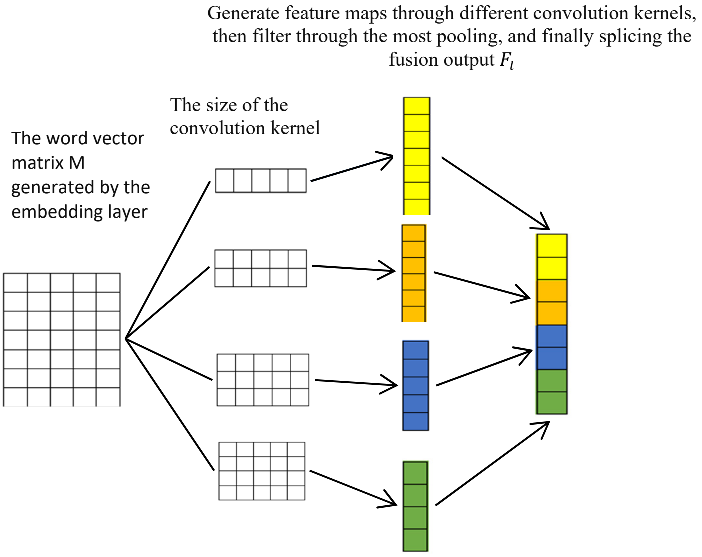
Generate feature maps through different convolution kernels, then filter through the most pooling, and finally splicing the fusion output F_l
The size of the convolution kernel
The word vector matrix M generated by the embedding layer

Figure 2. CNN network

3.3. BiLSTM layer

Although *LSTM* has improved *RNN*, the network model can consider the one-way text information of the sentence. Since the text is composed of context information, only considering the text information on the left side without considering the text information on the right side will cause information loss and affect the text. On this basis, an improved model BiLSTM based on long short-term memory network is proposed. BiLSTM is proposed for the problem that LSTM cannot consider the global structure of sentences. Therefore, the SMNN neural network model adopts a bidirectional LSTM network structure to capture the context information of each word. The bidirectional LSTM obtains the output s_t and l_t of the hidden state at time *t* through a forward LSTM and a reverse LSTM, and then Splicing the input s_t of the hidden state of the forward LSTM and the output$l_t$of the hidden state of the reverse LSTM to obtain the output h_t of the final hidden state of the bidirectional LSTM,h_t is the output of the hidden state of the bidirectional LSTM corresponding to the time t, that is, the corresponding word at time t vector representation. Taking the forward long short-term memory network as an example, the LSTM calculation process is as follows:

$$i_t = \sigma(W_i \cdot [s_{t-1}, e_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o \cdot [s_{t-1}, e_t] + b_o) \quad (6)$$

$$f_t = \sigma(W_f \cdot [s_{t-1}, e_t] + b_f) \quad (7)$$

$$c_t = f_t * c_{t-1} + i_t * tanh(W_c \cdot [s_{t-1}, e_t] + b_c) \quad (8)$$

$$s_t = o_t * tanh(c_t) \quad (9)$$

Among them,i_t represents the input gate,o_t represents the output gate,f_t represents the forget gate, s_{t-1} represents the word representation vector at the t-th position of the word vector generated by the pre-trained language model, and σ represents the sigmoid activation function , *W* is the weight matrix involved in the operation, *b* represents the offset, c_t represents the output of the hidden layer, and finally the output of the hidden layer and the output of the output gate jointly determine the output element s_t.

Although the long short-term memory network solves the gradient disappearance and gradient explosion problems of RNN to a certain extent, and can capture the context information of the sequence well, but because the long short-term memory network introduces the forgetting gate, the LSTM propagation process will generate memory. Missing is the loss of some important data, which affects the experimental effect. In order to prevent semantic loss during the training process, the SMNN model uses the word embedding model based on the attention mechanism to initialize and generate the word vector e_t, and then the vector e_t output by the embedding layer is input into BiLSTM to output the feature vector sl_t, and finally the output feature vector after splicing is spliced. sl_t is then output through maximum pooling. On the one hand, this model stacking method increases the depth of the network and helps to improve the performance and training efficiency of the model. On the other hand, it helps to deeply capture text features and sentence structure. The connected structure helps avoid exploding and vanishing gradients during model training, as shown in Figure 3.

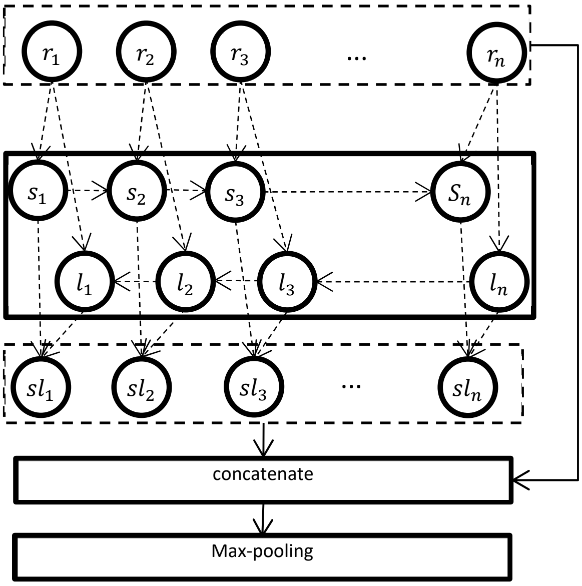

Figure 3. BiLSTM network

3.4. Fusion output layer

By inputting the global semantic feature representation generated by the self-attention mechanism obtained by the embedding layer into the global feature representation F1 obtained by the max pooling layer, the feature representation obtained by the CNN layer and the feature representation generated by the BiLSTM layer are spliced in dimension. The global feature representation F, such a global representation has rich text feature representation. In order to speed up the training, the tanh activation function is introduced, and random deactivation is introduced to allow the neural units in the network to be discarded from the network in a certain proportion. The purpose of this is to Improve the generalization ability of the model to prevent overfitting during model training, then input it into the full connection, and finally input it into the softmax to get the final prediction probability *P*.

$$P=softmax(W_sF + b_s) \tag{10}$$

where W_s represents the trainable weight,$b_s$represents the bias term, and finally the cross-entropy loss function is used as the target function for text classification training.

$$H(p,q)=\sum_{i\in N}\sum_{j\in C} q_i^j logp_i^j \tag{11}$$

Among them, *N* is the number of training samples, *C* is the number of categories, *q* is the true label of the sample, and the one-hot encoding used.

4. Analysis of results

4.1.Experimental dataset

In order to verify the superiority of the SMNN network model in text classification, related experiments are carried out on two public datasets, and the relevant information of the two datasets is shown in Table 1. Among them, length represents the average length of the data, class represents the number of classifications, train represents the number of samples in the training set, dev is the number of samples in the validation set, and test refers to the number of samples in the test set:

Table 1. Data set information statistics

Data	train	dev	test	class
THUCNews	180000	10000	10000	10

THUNews_Title dataset: THUCNews is generated by filtering and filtering the historical data of the Sina News RSS subscription channel from 2005 to 2011, with a total of 74 records. The dataset is short text data of multi-text classification. This article extracts 10 categories from the THUNews dataset for training, which are finance, realty, stocks, education, science, society, politics, sports, game and entertainment. The total number of data sets is 200,000, the training set is 180,000, and the validation set and test set are 10,000 each.

4.2.Experimental parameter settings

Weight initialization: The model weights need to be initialized during model training. The SMNN model is initialized by random sampling from the uniform distribution of [-1,1].

Training hyperparameters: The word embedding and self-attention mechanism used by SMNN are used as word embeddings. The dimensions of the output embedded word vector and word vector are both 300, and the parameters of the embedding layer will be updated as the word vector is generated. Four one-dimensional convolution kernels of different sizes are used in the CNN layer for vertical convolution. The size of the convolution kernel is 3, 4, 5 and 6, and the number of each convolution kernel is set to 256. The number of hidden units in BiLSTM is set to 128, the number of hidden layers in BiLSTM is set to 2, and a dropout mechanism is introduced to prevent overfitting of the model, dropout is equal to 0.2, the minimum batch size during model training is 128, and the training of the model The parameters are optimized using the Adam optimizer with a learning rate of 0.003. The relevant parameters of the SMNN model are shown in Table 2.

Table 2. SMNN network parameter settings

hyperparameter name	**values**
word vector dimensions	300
Kernel size	（1,2,3,4）
Number of kernel	256
hidden units	128
epochs	20
optimizer	Adam
learning rate	0.003
dropout	0.4
Pad_size	100
num_layers	2

4.3.Experimental results and analysis

4.3.1.Comparative Test

The SMNN model is compared with the widely used classification models, which mainly include TextCNN, TextRCNN, DPCNN, Att-BiLSTM and Transformer.

- TextCNN: Kim proposed to use multiple convolution kernels of different sizes to do vertical convolution in the CNN network to extract the n-gram features of the text, process the data through the activation function (relu) to speed up the model training, and finally pass the maximum pooling. to extract the most important features of the text through softmax classification.
- TextRNN: Liu et al. proposed a recurrent neural network structure for text classification. This structure mainly uses the output of the hidden layer of the text in the last time step of LSTM as the feature representation of the global text semantics, and finally passes the classifier (softmax). Classification.
- TextRCNN: A new network structure proposed by Lai et al. is used for text classification. This network structure fully draws on the network structure of TextRNN. Different from TextRNN, the maximum pool is added after the feature representation of the text is learned through the recurrent neural network. to extract the salient features of the text.
- DPCNN: Johnson et al. proposed a new pyramid-like network structure for text classification, which used increasing the depth of the network to improve the performance of DPCNN.
- Att-BiLSTM: Zhou et al. proposed to capture the global semantic features of sentences by combining attention mechanism and bidirectional long-term and short-term memory network, and using attention mechanism to assign different features to different words to capture important semantic information of texts.
- Transformer: Transformer was proposed by Vaswani et al. in machine translation, which consists of an encoder and a decoder. In the text classification task, the encoder is used to obtain long-distance features of the text.

4.3.2.Model comparison analysis

This paper uses the accuracy rate (acc), precision rate (Precision), recall rate (Recall), and F1 value (F1-socre) of the model to evaluate the proposed SMNN model. The calculation formulas of these four indicators are as follows:

$$Acc=\frac{TP+TN}{TP+TN+FP+FP} \tag{12}$$

$$Precision=\frac{TP}{TP+FP} \tag{13}$$

$$Recall=\frac{TP}{TP+FN} \tag{14}$$

$$F1=\frac{2Precisoin\times Recall}{precison+Recall} \tag{15}$$

Among them, the total number of correct samples predicted by TP, the total number of wrong samples predicted by FP, the total number of samples for actual text classification by TP+FP, the total number of samples that should be classified by TP+FN, and the comprehensive index F1 is obtained by comprehensively considering the precision rate and the recall rate.

The classification results of this experiment are shown in Figure 4. Through the comparative analysis of the results, it can be seen that the SMNN model has the best classification effect on sports and education, and their F1 values exceed 95%. For properties, games are the next most effective category, but their category F1 scores are also over 92%. Then there's entertainment, politics and society, which have F1 averages between 91% and 92%. For finance, stocks and science were less effective, with F1 values ranging from 86% to 89%. Overall, the SMNN model performs well in various classifications, indicating that the SMNN model has a superior classification effect on texts and can accurately achieve text classification.

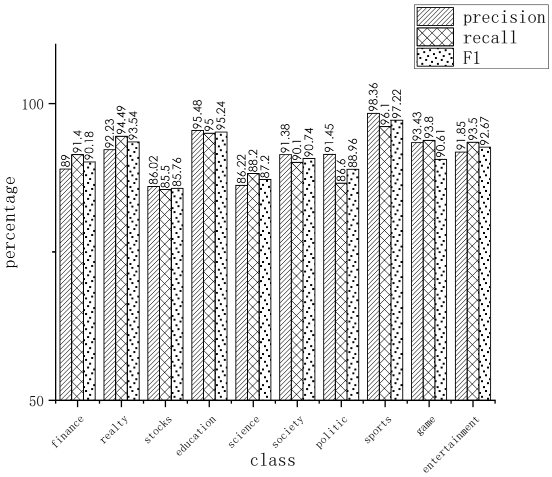

Figure 4. Comparison of evaluation indicators for different classifications of the SMNN model

As shown in Figure 5, the SMNN model has a very superior performance on the Sina news dataset. Its accuracy, precision, recall, and F1 are 91.51%, 91.54%, 91.51% and 91.51% respectively. In the first 6 experiments, the best experimental result is the RCNN model. Compared with the RCNN model, the SMNN model has achieved 1.05%, 1.15%, 1.14%, and 1.12% in the accuracy rate, precision rate, recall rate and F1 respectively. Through the comparison of the accuracy, precision, recall and F1 score of the model, it is proved that the SMNN model extracts text signs with convolutional neural network and recurrent neural network respectively and improves the performance of the model. Compared with the traditional text classification model, the classification effect of the SMNN model is significantly improved.

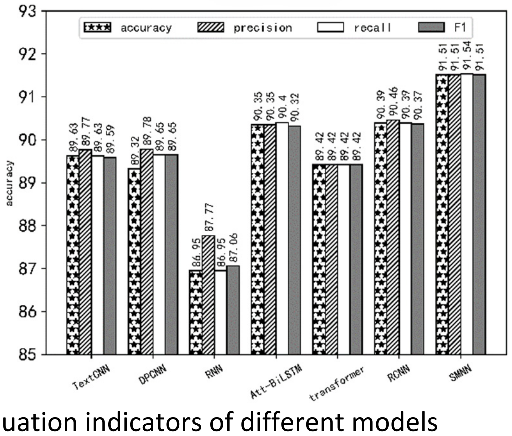

Figure 5. Comparison of evaluation indicators of different models

As shown in Figure 6, the accuracy comparison chart of RCNN, Att-BLSTM and SMNN training process is shown. The abscissa is the number of iterations of the dataset (unit is epoch, 100 batches, and each batch has 128 data), and the ordinate is the accuracy of validation set. with the increase of the training data set, the accuracy of the model changes in the validation set. the RCNN model, the Att-BiLSTM model and the SMNN model with the best performance among the above models are selected, and the results show that the SMNN model is on the validation set outperforms the RCNN and ATT-BiLSTM models.

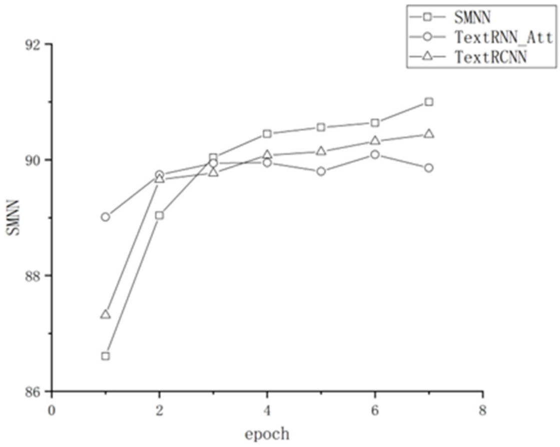

Figure 6. The accuracy comparison of model training process

4.3.3.Comparative Test

In order to verify the influence of different modules of the SMNN model on the classification effect of the model and further prove the effectiveness of the model, an ablation experiment is designed on this basis. On the basis of SMNN as the original model, three groups of experiments were divided for comparison. The first group is Word2vec-Attentation, which inputs the feature representation of the output of the word embedding model based on the attention mechanism into the fusion output layer for text classification, that is, directly removes the CNN layer and BiLSTM layer on the basis of the original model; The second group of Word2vec-Attentation-CNN is to output the feature representation of the output of the embedding layer to the CNN layer, extract text features through convolution kernels of different sizes, and finally input them to the fusion output layer for text classification; the third group is Word2vec -Attentation-BLSTM, which outputs the feature representation generated based on the attention mechanism and Word2vec to the BLSTM layer, obtains the global feature representation of the text through the long short-term memory network, and finally classifies it through the fusion output layer. The experimental results of ablation are listed in Table 3.

Table 3. Ablation experiment results

Network of fusion	precision(%)	recall(%)	F1(%)	accuracy(%)
Word2vec-Attentation	88.88	88.83	88.84	88.83
Word2vec-Attentation-CNN	90.95	90.92	90.92	90.92
Word2vec-Attentation-BLSTM	87.58	87.15	87.19	87.15
SMNN	91.51	91.54	91.51	91.51

It can be seen from Table 3 that the classification effects of the model Word2vec-Attentation, the model Word2vec-Attentation-CNN and the model Word2vec-AttentationBLSTM are far less than the classification effect after the model fusion. As shown in Figure 7, the accuracy of the model training process is compared. It can be seen from the figure that the SMNN model is higher than other decomposition models in terms of data convergence speed and accuracy. The SMNN model can fuse the global semantic features of the text with the local semantics at multiple granularities, and has stronger semantic capture and information extraction capabilities.

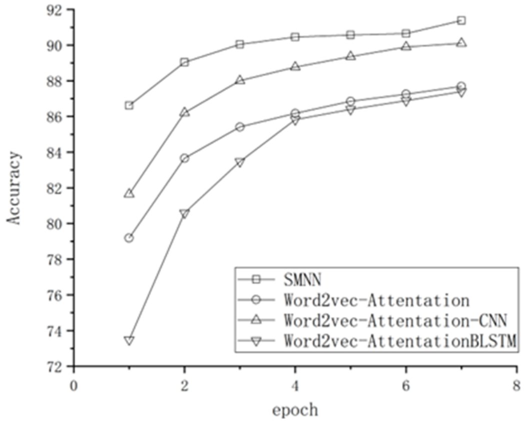

Figure 7. Ablation experiment process comparison

5. Conclusion

The experimental results show that the SMNN text classification model has certain advantages compared with the traditional text classification model. The word embedding model based on the self-attention mechanism obtains the global representation of the text, uses CNN to extract the local semantic features of the text at multiple granularities through different convolution kernel sizes, and uses the BiLSTM with skip connections and the pooling layer to obtain the global text of the text. Semantic features. The SMNN model has stronger feature extraction capabilities than a single attention-based word embedding model (Word2vec-Attentation), CNN, BiLSTM, and an attention-based word embedding model and a single CNN and BiLSTM combined model, which can Sufficient global and local features of the text can achieve better classification results in each classification, and the self-attention mechanism can improve the performance of the model in the process of combining with other neural network models. In the following research, we will explore how to combine the self-attention mechanism is combined with other deep network structures for text classification.

6. References

[1] MARON M E, KUHNS J L. On relevance，probabilistic indexing and information retrieval[J]. Journal of the ACM, 1960, 7(3):216-244.

[2] HINTON G E，SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science,2006,313(5786): 504-507.

[3] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv.preprint.arXiv:1404.2188,2014.

[4] KIM Y, et al. Convolutional neural networks for sentence classification [J]. arXiv.preprint.arXiv:1408.5882,2014.

[5] JOHNSON R， ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. DOI: 10.18653/V1/P17-1052.

[6] CONNEAU A，SCHWENK H，BARRAULT L，et al. Very deep convolutional networks for text classification [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017.DOI: 18653/V1/E17-1104.

[7] JORDAN M I. A parallel distributed processing approach[J]. Advances in Psychology ， 1997,121:471-495.

[8] ZHANG Y，LIU Q，SONG L F. Sentence-state LSTM for text representation [C]// Proceeding of the 56th Annual Meeting of the. Association for Computational Linguistics.2018.DOI:10.18653/V1/P18-1030.

[9] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473，2014.

[10] LUONG M T，PHAM H，MANNING C D. Effective approaches to attention-based neural machine translation [J]. arXiv preprint.arXiv:1508.04025，2015.

[11] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[12] Jia Hongyu, Wang Yuhan, Cong Riqing, et al. NEURAL NETWORK TEXT CLASSIFICATION ALGORITHM COMBINING SELF-ATTENTION .MECHANISM[J].COMPUTER APPLICATIONS AND SOFTWARE. 2020,37(2): 200-206.

[13] Xinqiang Li,Weina Niu,Xiaosong Zhang, et al. Improving Performance of Log Anomaly Detection With Semantic and Time Features Based on BiLSTM-Attention[C]//.Proceedings of 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT 2021).,2021:697-702.DOI:10.26914/c.cnkihy.2021.065498.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.

[15] PENNINGTON J, SOCHER R, MANNINGC D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.(EMNLP). 2014;1532-1543.

[16] Radford, et al. Language models are unsupervised multitask learners[J]. Open AI Blog,2019,1(8):9.

[17] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies. 2019:4171-4186.

[18] LI K Y, CHEN Y, NIU Sz. Social E-commerce Text Classification Algorithm Based on BERT[J/OL]. Computer Science, 2021,48(2):87-92.

[19] RASMY L, XIANG Y, XIE Z Q, et al. Med-BERT : pretrained contextualized embedding on large-scale structured electronic health records for disease prediction[J].NPJ Digital Medicine,2021,4(1): 1-13.

[20] LAN Z, CHEN M, GOODMAN S, et al. Albert : A lite bert for self-supervised learning of language representations[C]// Proceedings of the 8th International Conference on Learning Representations. ICLR,2020:1-17.

[21] Zhang Zhong-lin, Li Lin-chuan, Zhu Xiang-qi, et al. Aspect sentiment analysis combining ON-LSTM and self-attention mechanism[J]. Journal of Chinese Computer Systems,2020,41(9):1839-1844.

[22] SHI Zhenjie, DONG Zhaowei, PANG Chaoyi, et al. Sentiment analysis of e-commerce reviews based on BERT-CNN[J]. INTELLIGENT COMPUTER AND APPLICATIONS,2020,10(02):7-11.

[23] Liu Jingru, Song Yang，Jia Rui，et al．A BiLSTM-CRF Model for Protected Health Information in Chinese[J]．Data Analysis and Knowledge Discovery，2020，4 (10): 124-133.