DOI: 10.31534/engmod.2022.2.ri.01d
Original scientific paper
Received: 20.07.2021.

Multi-model Fusion Attention Network for News Text Classification

Zhengpeng Li(1), **Jiansheng Wu**(1*), **Jiawei Miao**(1), **Xinmiao Yu**(1), **Shuaibo Li**(1)

(1) *School of Computer and Software, University of Science and Technology Liaoning, Anshan 114044, CHINA*

(*) *corresponding author, e-mail: ssewu@163.com*

SUMMARY

At present, the classification prediction task based on news content or news headline has the problems of inaccurate classification and attention deviation. In this paper, a multi-model fusion attention network for news text classification (MFAN) is proposed to train news content and news titles in parallel. Firstly, the multi-head attention mechanism is used to obtain the category information of news content through a dynamic word vector, focusing on the semantic information that significantly influences the downstream classification task. Secondly, the semantic information of news headlines is obtained by using the improved version of the long-short-term memory network, and the attention is focused on the words that have a great influence on the final results, which improves the effectiveness of model classification. Finally, the classification fusion module fuses the probability scores of news text and news headlines in proportion to improve the accuracy of text classification. The experimental test on the Tenth China Software cup dataset shows that the F1 - Score index of the MFAN model reaches 97.789 %. The experimental results show that the MFAN model can effectively and accurately predict the classification of news texts.

KEYWORDS: *multi-model fusion; attentional mechanism; text classification; neural network.*

1. INTRODUCTION

With the rapid development of the news industry, the traditional paper news text has tended to electronic and Datamation text. The number of daily news readings has exceeded one million, and the classification of news text is becoming more and more important. With the development of science and technology, the method of text classification is also changing, and people need a fast and efficient classification retrieval method to meet people's fast-paced lifestyles [1]. The application field of text classification is also expanding, playing a key role in the field of e-commerce, public opinion analysis, and other application fields. Text categorization, being the process of assigning predefined labels to text, is a fundamental natural language task. The traditional shallow learning model acquires the sample features by manual methods and a machine learning algorithm is used to classify the text. Therefore, the effectiveness of shallow learning methods is limited by feature extraction to a large extent [2]. Deep learning integrates feature engineering directly into the output by learning a set of nonlinear transformations, thus integrating feature engineering into the model fitting process.

From the late twentieth century to the early twenty-first century, text classification models are mostly based on shallow learning or statistical models: for example, the Naive Bayesian Model (NBM) suggested by Maron and Kuhns [3], K-Nearest Neighbor (KNN) proposed by HM Takahashi [4], and Support Vector Machine (SVM) [5-6] proposed by Vapnik et al. In machine learning, people often use the Vector Space Model (VSM) [7] combined with machine learning classifier to achieve text classification. At the early stage of the 21st century, artificial neural networks (ANNs) evolve into neural network structures with constantly improving Learning ability, being summarized as Deep Learning (DL) [8]. The wide use of the deep learning model has led to the technological revolution of the text classification model. The use of deep learning greatly reduces the workload of manual rules and pays more attention to the representation information of text semantics [9]. Text classification was introduced into the news field in China around 2013, mainly reflected in the news classification processing by Xinhua News Agency and other media using classification technology.

With the rapid development of deep learning, the use of pre-training models and attention mechanisms in natural language processing entered a new era [10]. Existing news text classification models based on deep learning train only content texts, news title [11] or news title and content texts together for training, these three training methods all ignore the correlation between the news title and the semantic level of the news content, which affects the accuracy of news category judgment. The spread of fraudulent news headlines is also a serious problem, and identifying the classification of fraudulent news headlines requires a lot of manual reviews. However, a manual review of a large number of fraudulent news titles is actually not feasible [12-13]. Therefore, the text classification model should not only learn the semantic features of news text but also combine the semantic information of news headlines, paying attention to the semantic association information between headlines and text.

News text is mostly report description text, the data sample length is large with many data types and uneven sample numbers. Due to the imbalanced phenomenon of news text data samples, the existing text classification models are not accurate enough to recognize news text classification tasks, and the models do not have a better generalization ability and scalability. Given the above problems in the process of news text classification, such as the unbalanced number of data samples and the insufficient attention of the model to semantic level information, this article makes the following improvement contributions:

- In this paper, a multi-model fusion news text classification method (MFAN), combining attention mechanism, is proposed. Considering the keyword vector requirement of the downstream classification task, the model uses a multi-head attention mechanism to obtain semantic information of news content.
- MFAN model proposes that the method of learning news content and titles in parallel can effectively improve the classification effectiveness of the model. Considering that news headlines are short texts, the MFAN model uses a cyclic neural network with gated units to obtain semantic information of news titles. This paper designs a classification fusion module to balance the attention score of news content and news headlines.
- In this paper, a large number of experiments are carried out on the proposed algorithm using the Tenth China Software Cup dataset. The experimental results show that the MFAN model is superior to most current classification model algorithms.

2. RELATED WORK

2.1 LONG SHORT-TERM MEMORY

The Long Short-term Memory (LSTM) model proposed by Hochreiter et al [14], is one of the Recurrent Neural Network (RNN) models used most frequently in recent years. LSTM model can effectively solve the gradient descent and gradient vanishing problems in conventional cyclic neural networks [15]. Compared with other cyclic neural network models, the biggest innovation of the LSTM model is the introduction of control gates classified into three types: InPut Gate, OutPut Gate, and ForGet Gate [14-15], as shown in Figure 1. The input gate learning of the LSTM model specifies the time when the activation signal is passed into the storage unit, the output gate learns the time when the activation signal is sent out of the storage unit, and the forgetting gate learns when the storage unit at the previous moment is passed into the storage unit at the next moment.

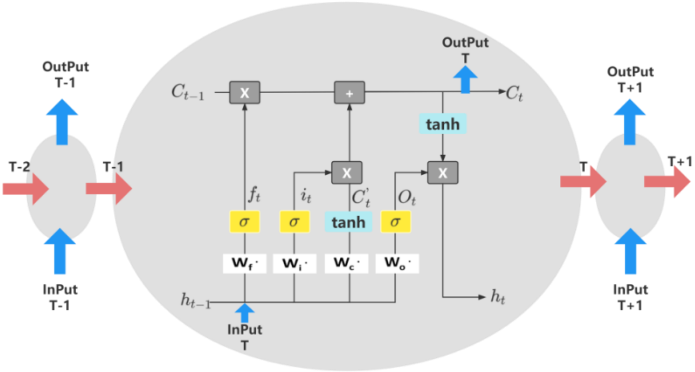

Fig. 1 *LSTM Model Structure*

The first step is to calculate what information is discarded in the ForGet Gate (f_t). h_{t-1} represents the output of the transition point state of the previous unit in the LSTM model, x_t represents the model data input in the current state, and W and b are both learnable and trainable weight matrices in the unit structure:

$$f_t=\sigma(W_f\cdot[\ h_{t-1},x_t\ ]\ +\ b_f\ ). \tag{1}$$

The second step selects the pre-updated data to be mapped to the interval ($0, 1$). The *tanh* layer is the *tanh* activation function. C'_t represents the new vector, which is the weight ratio of the current calculated value to be updated to the memory unit:

$$i_t\ =\ \sigma(W_i\cdot[\ h_{t-1},x_t\ ]\ +\ b_i\ ), \tag{2}$$

$$C'_t\ =\ tanh\ (W_c\cdot[\ h_{t-1},x_t\ ]\ +\ b_c\ ). \tag{3}$$

In the third step, the old state C_{t-1} in the unit structure is updated to the new memory state C_t, where ⨀ represents the multiplication of matrix elements:

$$C_t\ =\ f_t\odot C_{t-1}+\ i_t\odot C'_t\ . \tag{4}$$

In the last step, the limit of the OutPut Gate on the output unit is calculated, and the hidden state h_t obtained:

$$o_t\ =\ \sigma(W_0[\ h_{t-1},x_t\ ]\ +\ b_a), \tag{5}$$

$$h_t = o_t \odot tanh\,(C_t\,). \tag{6}$$

2.2 ATTENTION MECHANISM

The attention mechanism is also the product of the rapid development of deep learning, widely used in Image Processing, Natural Language Processing (NLP), Data Prediction, and other fields. The essence of the attention mechanism is to calculate the weight coefficient corresponding to the source sequence (Query) to the target series (key-value) and then focus on the key points of the source sequence through the weight coefficient. The structure of the attention mechanism model is shown in Figure 2.

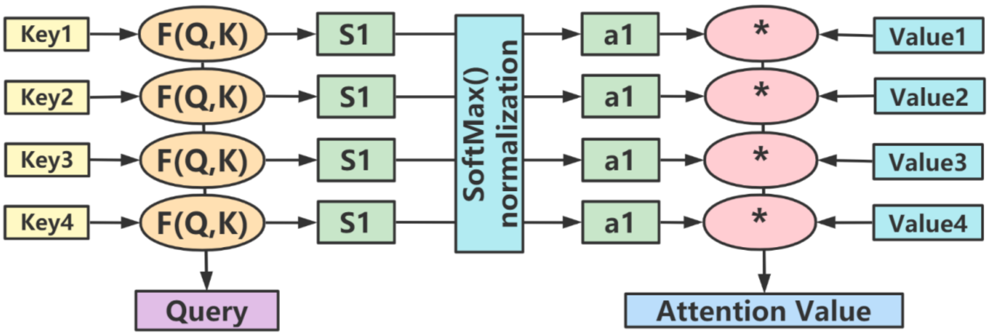

Fig. 2 *Attention Mechanism Structure*

The main role of the Attention mechanism is to capture the proportion of input semantic information that is weighted and to control the scores on different tasks. If the input is important to the task execution, the attention mechanism will strengthen the weight of this important information and focus the input with a high weight. On the contrary, if the input information has little or no influence on the task execution, the weight value obtained will be appropriately reduced, even to zero [16]. For the same input information, the weight value obtained in different tasks may also differ. This attention mechanism of constant regard to task requirements greatly improves the generalization ability of the network model.

2.3 BERT MODEL

In 2018, the BERT (Bidirectional Encoder Repres-Representation from Transformers) pretraining model was proposed by Devlin J, Chang M.W., Google. BERT is a large multitask language model [17]. Word embeddings is a new concept in the field of natural language processing. As one of the most common methods of text feature representation, word vector as input feature has been widely used in natural language processing tasks [18]. BERT model is also known as Encoder of bi-directional Transformer as shown in Figure 3. Decoders are not used because they cannot obtain the information to be predicted. Compared with other network models, the major difference between the BERT model and other network models is that the pre-train part of the BERT model uses the Masked Learning Model and the Next Sentence Prediction. Such two methods capture Representation at the word level and the sentence level, respectively, which are also the main tasks of the pre-training of the BERT model: to produce random mask LM and to predict the next semantic information (NSP). Based on the excellent scalability and fine-tuning characteristics of the BERT model, NLP researchers mostly do not need to start training their own models from scratch. Based on the self-attention mechanism in Transformer [19], BERT is allowed to simulate many downstream tasks by changing the appropriate input and output, as shown in Figure 3.

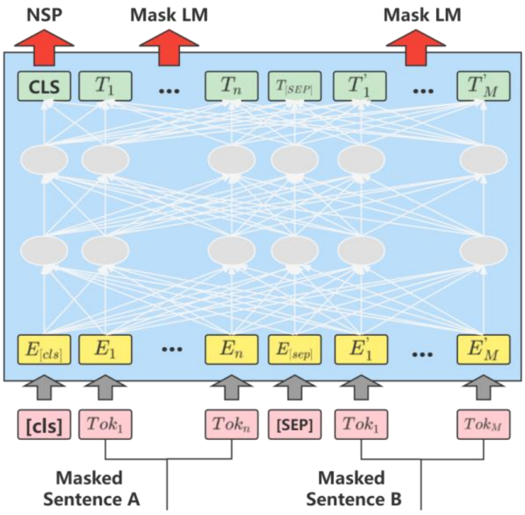

Fig. 3 *BERT on different tasks*

3. MFAN MODEL

This section introduces the multi-model fusion attention network for news text classification (MFAN). The structure of the model, as shown in Figure 4, consists of three parts: content feature extraction module, title feature extraction module and Classification fusion module.

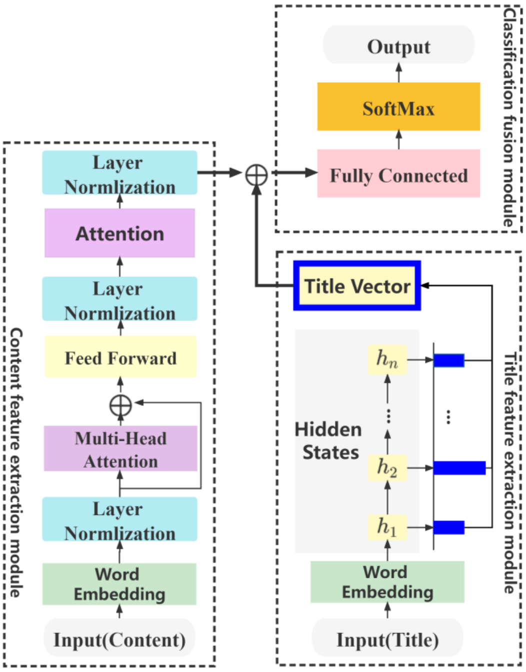

Fig. 4 *MFAN Model Structure*

3.1 CONTENT FEATURE EXTRACTION MODULE

The content texts feature extraction module is composed of the BERT network model and Attention, as shown in Figure 5. Firstly, the content texts is preprocessed by truncating the text to preserve the appropriate number of text words. The size of the input after truncation is [batch_size, max_segment, max_len], where batch_size is the size of a batch, max_segment is the maximum number of sentences after truncation, and max_len is the maximum length of each sentence. The truncated "sentence" is transmitted to the BETR model to obtain the sentence vector. Principle of text truncation: Truncate the text into several paragraphs according to the maximum length max_len. If the number of paragraphs exceeds the maximum number of paragraphs max_segment, only characters with max_len at the beginning and max_len at the end of the text are taken, and the characters with max_len at the end of the text are discarded.

Principle of text abandonment: Generally, in a long sentence, the beginning and the end often contain sufficient semantics, while the semantics in the middle part is relatively weak. Abandoning the weak semantic information can effectively improve the efficiency of the network model. The first truncated max_len length character is converted into word vector *P1*, and the last truncated max_len length character is converted into word vector *P2*, and the word vector is expressed as P: $\{w_1^d, w_2^d, w_3^d, \dots, w_n^d\}$.

For the better integration of the information of each paragraph, it is necessary to capture the connection between each paragraph in the truncated long text. After the multi-head attention layer, the attribute value of each group of characters class label (cls_n) is extracted and transmitted to the Attention layer. The vector representation of news text is obtained through the attention layer.

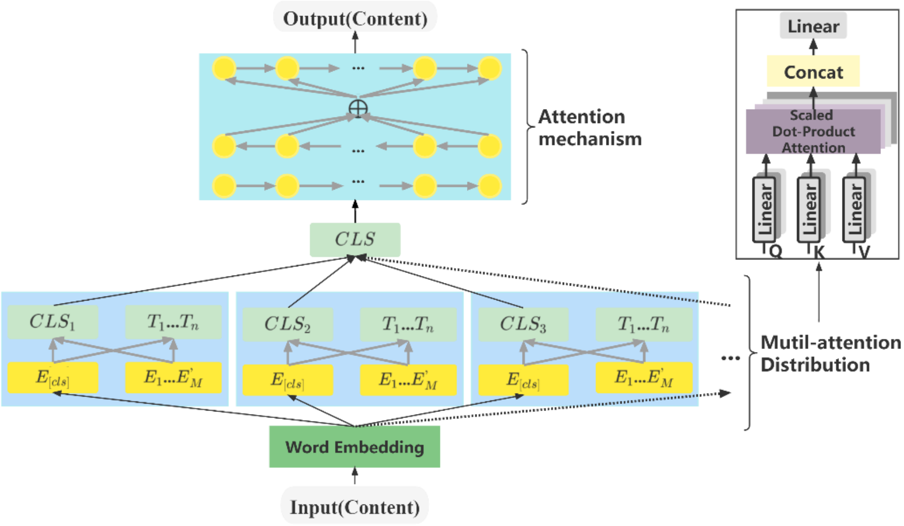

Fig. 5 *Model Structure of News ConTexts*

$$K_i = W \times cls_i + b \ , \tag{7}$$

$$f \ Q,K_i \ = Q^T K_i, \tag{8}$$

$$a_i = softmax(f \ Q,K_i \ = \frac{exp \ f \ Q,K_i}{\sum_{j=1}^{n} exp \ f \ Q,K_j}, \tag{9}$$

$$Attention(Q,K,V)=SoftMax(\frac{(QK^{T})}{\sqrt{(d_k)}})V, \tag{10}$$

$$Space_i = Attention\ \ QW^{Q_i}\ ,KW^{K_i}\ ,VW^{V_i}\ \ , \tag{11}$$

$$O_c = MultiHead = Concat\ \ Space_1,Space_2,\ldots,Space_i\ \ W^{dr}, \tag{12}$$

An additional scaling factor $\sqrt{d_k}$ is introduce and the matrix W^{dr} is used for multi-space fusion. O_c represents the output vector of the news content, i represents the *i-th* of text sequences of the input news content, W represents the weight of the Attention layer, b represents the bias in the Attention layer, and Q is the query vector of K. *W, b, Q* are the trainable parameters.

3.2 TITLE FEATURE EXTRACTION MODULE

In this article, news title data is classified and trained on the basis of LSTM+ATT module. The structure of the news text title network model is divided into three parts: the input layer, the LSTM layer and the subsequent processing part. The specific model structure is shown in Figure 6.

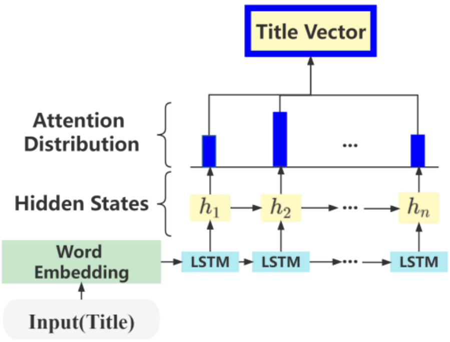

Fig. 6 *Title feature extraction module*

In the first part, the input layer consists of a model input module (news title) and a word vector (word embedding) layer. The input module unifies the news title to the same length and transmits it to the Word Embedding layer. The Word Embedding is used to convert each Chinese character of an input news title into a *200*-dimensional word vector and pass it to the LSTM Layer. The input method of word embedding is an improvement of distributed representation proposed by Hinton [20]. Word embedding is proposed to effectively solve the problem of input dimension explosion and overcome the one-hot semantic information that cannot reflect the correlation between words.

$$w_t = \sum_{1}^{n} L_n \times d \tag{13}$$

w_t is the word vector representation of all news headlines, $L = \{L_1, L_2, \ldots, L_n\}$ is the input of n news titles, and the maximum input character number max_titlelen is set for each news title input into the LSTM model. d is the dimension of the news title vector.

The second part is the core part of the News headlines network model. The LSTM+att module sequence is used to effectively learn the context information of News headlines and the learned feature information and news content texts will be spliced and passed to the classifier to realize the integration of the semantic relevance information of News headlines and News texts.

3.3 CLASSIFICATION FUSION MODULE

In this section, the output matrix of the news title is fused with the content texts matrix introduced above, and input to the full connection layer for subsequent classification processing. In order to reduce the occurrence of overfitting and ensure the regularization effect, the hidden nerves in the network are randomly deleted by dropout to keep the input and output of the network model unchanged. The content texts data processed by the Bert model and attention mechanism and the news title data processed by the LSTM model are fused as the input of the full connection layer, and then classified and judged by the softmax layer. The overall structure is shown in Figure 7.

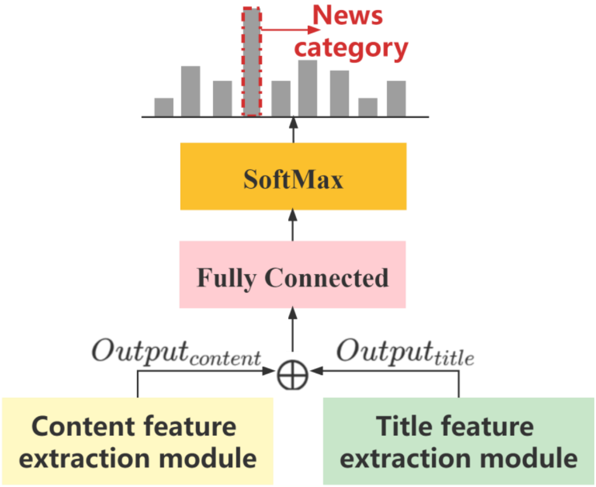

Fig. 7 *Classification fusion module structure*

The Softmax layer uses the Softmax regression model to learn the characteristics of the information transmitted from the LSTM hidden layer, calculate the probability that the pre-classified news text data belongs to categories, and transmit it to the output layer. Finally, the predicted category of news text to be classified is given, and “OutPut” is the OutPut of the whole model:

$$News = Output\_Content \oplus Output\_title, \tag{14}$$

$$OutPut = soft\max(News). \tag{15}$$

4. EXPERIMENTS

4.1 DATASETS

This article uses the data sets provided by the Tenth China Software Cup dataset. The data set includes news text information in *9* fields of finance and economics, real estate, education, science and technology, military, automobile, sports, games, and entertainment. It includes

14,641 pieces of data. The number of news text messages with labels in each category is shown in Figure 8. After deleting the noisy data from the labeled *14641* data sets, *14405* news texts were left. *100* pieces of each category randomly selected, and a total of *900* pieces of data were used as the test.

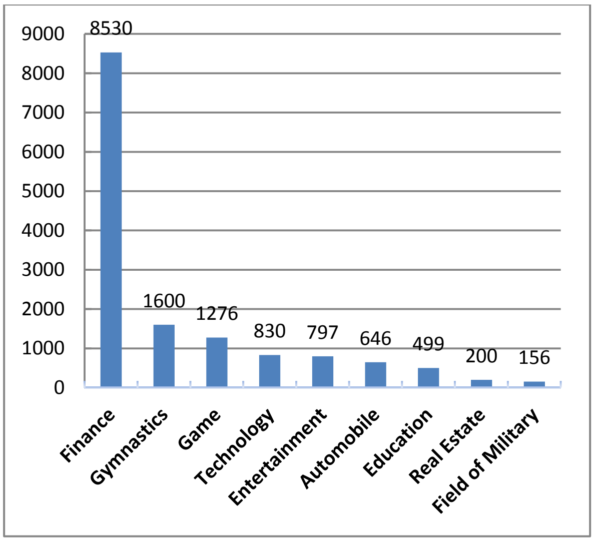

Fig. 8. *News text dataset comparison*

4.2 DATA PREPROCESSING

The Tenth China Software Cup dataset is in the format of Chinese, so it is not possible to use space and other delimiters for word segmentation. The jieba Chinese word segmentation tool is used for word segmentation. The characters of the news content texts are intercepted in accordance with the principle of text truncation and used as the data input of the content feature extraction module. The dimension of word vector *200* is set and the shuffle function is used to randomly shuffle the data. Headlines enjoy the same dimension of word vectors and the same way of shuffling data.

4.3 BASELINES AND IMPLEMENTATION DETAILS

The experiment of this paper is carried out under the Ubuntu18.04 system. The programming language of the experiment is Python3.0, the development tool is PyCharm, and the deep learning framework is Tensorflow1.14. The model was trained on the GTX 2070TI(8G) and GTX 3060(12G).

In order to verify the MFAN model, the following baseline models were compared respectively:

- **LSTM** [14] is a cyclic neural network model, which uses embedding to obtain word vectors, and then carries out feature representation of news titles based on word vectors.
- **CNN+LSTM+ATT** [21] uses CNN and LSTM to extract text local information and context features. Multi-channel Attention mechanism (Attention) is used to extract the Attention score of CNN and LSTM outputs. The output information of the multi-channel attention mechanism is fused to effectively extract text features and focus attention on important words.

- **CNN+BiGRU+ATT** [22] uses word2vec to vectorize the initial text, and selects window values through experiment to form three channels. Then, CNN's strong learning ability is used to extract local features, bidirectional gated cyclic unit (BiGRU) is used to extract global contextual information, and attention layer and pooling layer are used to obtain and optimize important features. Finally, text classification is carried out.
- **BERT** [17] is a heavyweight model that uses Encoder in Transformer for model pre-training. The sample data used for downstream tasks is relatively small, so the data used for classification should be specially annotated. BERT pre-training method greatly accelerates the convergence rate of the model.
- **DCNN+BiGRU+BERT** [23] model uses the BERT method to train the language model of word semantic representation. The semantic vector is dynamically generated according to the context of the word and put into the DCNN-BigRU hybrid model. In this way, the semantic vector contains both local features of the text and contextual features of the text, which then improve the accuracy of text classification.

The parameter settings of the MFAN model are as follows:

- The dimension of word vector is dim=*200*, batch processing quantity batch=1, training round epoch=*15*, the maximum length of news document content data input maxlen=*512*, and the maximum length of news document title data input max_titlelen = *30*.
- Dropout = 0.2, Dropconnect [24] was adopted to prevent over-fitting in the process of model training. DropConnect works on the weight matrix between the hidden layers, changing each input weight attached to the node to zero with a probability of *1-p*.
- Using Adam Loss function [25] (Adaptive Moment Estimation), class_weight is used before the Loss function to solve the problem that Loss function pays insufficient attention to samples with fewer data due to uneven data sets.
- Initial learning rate *lr = le-5*. If there is no learning change in the two-round model, the learning rate will be reduced $lr_{new} = \frac{1}{2} lr_{previoustime}$.

4.4 EVALUATION MEASURES

Considering that the model task is a classification task, in order to evaluate the classification effectiveness of the news text classification model designed in this paper more comprehensively and objectively, *F1*-Scroe is used as the main evaluation index of this model. *F1*-Scroe is the harmonic mean of Precision (*P*) and Recall (*R*), and the larger the *F1*-score is, the better the classification effectiveness of the model is:

$$P = \frac{TP}{TP + FP}, \tag{16}$$

$$R = \frac{TP}{TP + FN}, \tag{17}$$

$$F1 - Score = \frac{2 \times P \times R}{P + R}, \tag{18}$$

TP: the actual sample is a positive one, and the classifier predicts it to be positive; *TN*: the actual sample is a negative one, and the classifier predicts it to be negative; *FP*: the actual sample is a negative one, and the classifier considers it to be a positive sample, so the prediction is wrong; *Fn*: the actual sample is a positive one, and the classifier considers it to be a negative sample, so the prediction is wrong.

5. EXPERIMENTAL RESULTS

In order to improve the authenticity and reliability of the experimental results in this paper, the average value of *10* experiments with the same parameters is used as the final result of the final model. The comparative experimental results are shown in Table 1. The MFAN model proposed in this paper trained the preprocessed news texts of class-weight respectively. After *10* epochs, loss tended to be stable, as shown in Figure 9.

Table 1 *Comparison of F1-Score prediction results of different algorithms on The Tenth China Software cup*

Models	*Datasets*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
LSTM	*Title*	*88.69*	*88.67*	*88.676*
LSTM	*Content*	*90.12*	*90.04*	*90.079*
CNN+LSTM+ATT	*Content*	*92.44*	*92.35*	*92.395*
CNN+LSTM+ATT	*Title+Content*	*93.18*	*93.11*	*93.145*
CNN+BiGRU+ATT	*Title+Content*	*94.11*	*94.18*	*94.145*
BERT	*Title+Content*	*95.87*	*95.94*	*95.872*
DCNN+BiGRU+BERT	*Title+Content*	*97.07*	*96.94*	*97.005*
MFAN(with all)	*Title+Content*	***97.85***	***97.78***	***97.789***

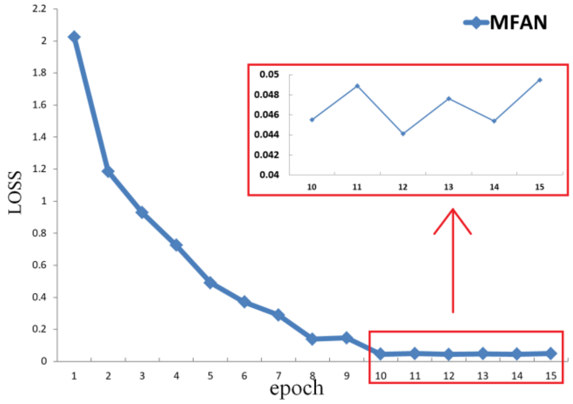

Fig. 9 *MFAN model loss curve*

Compared with the *F1*-score obtained by experimental models, the MFAN model proposed in this paper improved *9.211%* in *F1*-score compared with the LSTM model (the first line of Table 1) alone. Experimental results show that it is not effective to directly use word frequency-based text representation on short text data(news title), and pre-training with a large amount of semantic information can effectively improve the expression ability of the model, thus improving the accuracy of the downstream classification task.

In lines 2-3 in Table 1, it is indirectly verified that the main text covers most of the semantic information of the news. However, with the increase of the number of words in the text, the recurrent neural network model will have the problem of long-range dependence, and the earlier information recorded in the memory unit will be diluted over time step, so the dependence relationship with the earlier time step information cannot be established. With the summative statement of news, the model can easily blur or even incorrectly select the semantic information of news without extracting and learning the semantic information of the news title.

By observing the information in lines 4-7 in Table 1, the MFAN model can more easily combine the semantic association information between titles and content of news texts. News content is the main influencing factor of news text classification, and news headlines play a perfect supplementary role in news content training results. This method is similar to the learning process of feature maps at different scales in computer vision, which uses feature maps at different scales (Title+Context) to learn news categories for multiple times.

In order to verify that the MFAN model can effectively improve the classification effectiveness of the model by learning news content and title in parallel, we conducted an ablation experiment, as shown in Figure 10, comparing the feasibility of the with “Title feature Extraction model” and with “Content feature Extraction model”, respectively. When title feature extraction is not used, the model classification accuracy decreases by *0.93* points, and when content feature extraction is not used, the model classification accuracy decreases by *5.05* points. News content is rich in semantic information, and the MFAN model effectively extracts contextual semantic information by using a multi-attention mechanism. After the attention module, such differences in text information extraction are amplified in the final results. For short text titles, such as those of news, the LSTM+ATT module is used to learn the key information effectively, and the attention is focused on the words that have a great influence on the final results, which improves the effectiveness of the model classification, and verifies the feasibility of learning news content and news titles in parallel.

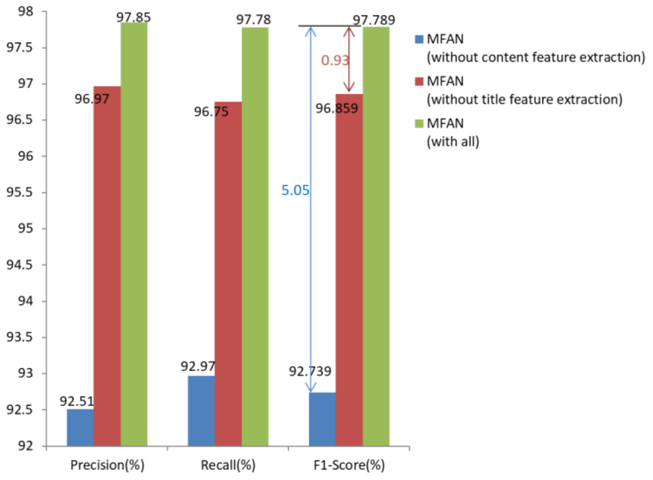

Fig. 10 *MFAN model ablation experiment*

6. CONCLUSIONS

In dealing with the multi-classification news text data with unbalanced data samples, the existing classification models study and analyze either only news text or only news headlines, or directly study and analyze news headlines and news text after splicing. These methods only learn semantic information and ignore the hierarchical semantic association information between news headlines and news text. In this paper, the MFAN model is proposed to learn and share parameters in parallel between news text and news headlines, which accelerates the convergence speed of the network model. The multi-attention mechanism is used to obtain the semantic information of news content, and the LSTM+att module is used to learn the key information of news titles. The attention is on the words that significantly influence the final results, which improves the effectiveness of model classification. In this paper, experiments are carried out on the Tenth China Software cup dataset. The research shows that the multi-module fusion news text classification method from this paper combined with the attention mechanism is far superior to the traditional classification model. Under the condition of basically unchanged detection speed, the model performs well in evaluation indexes, and the *F1* - Score reaches *97.789 %*. The future work will consider how to enhance the generalization ability and scalability of the MFAN model and apply it to other downstream tasks of natural language.

7. ACKNOWLEDGMENT

The research work was supported by National Natural Science Foundation of China (No.51774179) and Science and Technology Innovation Project of University of Science and Technology Liaoning (LKDYC202109).

8. REFERENCES

[1] J. Kim, S. Jang, E. Park, S. Choi, Text classification using capsules, *Neurocomputing*, Vol. 376, pp. 214-221, 2020. https://doi.org/10.1016/j.neucom.2019.10.033

[2] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P.S. Yu, L. He, A Survey on Text Classification: From Shallow to Deep Learning, *ACM Comput. Surv.*, Vol. 37, No. 4, Article 35, Publication date: July 2020, *Computer Science*, 2020.

https://doi.org/10.48550/arXiv.2008.00364

[3] M.E. Maron, J.L. Kuhns, On relevance, probabilistic indexing and information retrieval, *Journal of the ACM* (*JACM*), Vol. 7, No. 3, pp. 216-244, 1960.

https://doi.org/10.1145/321033.321035

[4] K. Hattori, M. Takahashi, A new edited k-nearest neighbor rule in the pattern classification problem, *Pattern Recognition*, Vol. 33, No. 3, pp. 521-528, 2000.

https://doi.org/10.1016/S0031-3203(99)00068-0

[5] C. Cortes, V.N. Vapnik, Support-vector networks, *Machine Learning*, Vol. 20, pp. 273-297, 1995. https://doi.org/10.1007/BF00994018

[6] V.N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995. https://doi.org/10.1007/978-1-4757-2440-0

[7] G. Salton, M.J. McGill, Introduction to Modem Information Retrieval, McGraw Hill Book Co., NewYork, l983.

[8] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, *Electronic Markets, The International Journal on Networked Business*, Vol. 31, No. 3, pp. 685-695, 2021. https://doi.org/10.1007/s12525-021-00475-2

[9] R.Y. Choi, A.S. Coyner, J. Kalpathy-Cramer, M.F. Chiang, J.P. Campbell, Introduction to Machine Learning, Neural Networks, and Deep Learning, *Translational Vision Science & Technology*, Vol. 9, No. 2, Article 14, 2020.

[10] J. Miao, J. Wu, Dialogue Model Based on the Improved Hierarchical Recurrent Attention Network, *International Journal for Engineering Modelling*, Vol. 34, No. 2, pp. 17-29, 2021.

https://doi.org/10.31534/engmod.2021.2.ri.02d

[11] X.Z. Dong, R. Song, Y. Hong, F.H. Zhu, Q.M. Zhu, News title classification based on multiple models, *Journal of Chinese Information Processing*, Vol. 32, No. 10, pp. 69-77, 2018.

[12] H.K. Liu, D.J. He, S. Chan, Fraudulent News Headline Detection with Attention Mechanism, *Computational Intelligence and Neuroscience*, Vol. 2021.

https://doi.org/10.1155/2021/6679661

[13] Seunghyun Yoon, Kunwoo Park, Minwoo Lee, Taegyun Kim, Meeyoung Cha, Kyomin Jung, Learning to Detect Incongruence in News Headline and Body Text via a Graph Neural Network, *IEEE Access*, Vol. 9, pp. 36195-36206, 2021.

https://doi.org/10.1109/ACCESS.2021.3062029

[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[15] K. Greff, R.K. Srivastava, J. Koutnik, B.R. Steunebrink, J. Schmidhuber, LSTM: A Search Space Odyssey, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 10, 2017. https://doi.org/10.1109/TNNLS.2016.2582924

[16] S. Hao, D-H. Lee, D. Zhao, Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system, *Transportation Research Part C: Emerging Technologies*, Vol. 107, No. 10, pp. 287-300, 2019.

https://doi.org/10.1016/j.trc.2019.08.005

[17] J. Devlin, M-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA: Association for Computational Linguistics, Vol. 1, pp. 4171-4186, 2019.

[18] J. Chen, J. Ma, X. Li, A Short Text Classification Method Combining Text Features of Pre-training Model, *Data Analysis and Knowledge Discovery*: 1-16 [2021-07-02].

http://kns.cnki.net/kcms/detail/10.1478.G2.20210625.1511.004.html

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

https://doi.org/10.48550/arXiv.1706.03762

[20] G.E. Hinton, Learning Distributed Representations of Concepts, Proceedings of the 8th Annual Conference of the Cognitive Science Society, 1986.

[21] J.B. Teng, W.W. Kong, Q.X. Tian et al, Multi-channel Attention Mechanism Text Classification Model Based on CNN and LSTM, *Computer Engineering and Applications*, Vol. 57, No. 23, pp. 154-162, 2021. DOI: 10.3778/j.issn.1002-8331.2104-0212

[22] N.T. Chen, J.H. Li, Y.Z. Man, Text Classification Based on Improved CNN-BiGRU-att Model, *Journal of Kunming University of Science and Technology* (Natural Sciences), Vol. 57, pp. 1-9, 2021. DOI: 10.16112/j.cnki.53-1223/n.2022.01.131

[23] H. Huang, X-Y. Jing, F. Wu, Y-F. Yao, X-Y. Zhang, X-W. Dong, DCNN-BiGRU text classification model based on BERT embedding, In 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS) IUCC/DSCI/SmartCNS, pp. 632-637, 2019.

https://doi.org/10.1109/IUCC/DSCI/SmartCNS.2019.00132

[24] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, R. Fergus, Regularization of neural networks using dropconnect, Proceedings of the 30th International Conference on Machine Learning, PMLR, Vol. 28, No. 3, pp. 1058-1066, 2013.

[25] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, Proceedings of the 3rd International Conferce for Learning Repressentations, ICLR, pp. 1-15, 2015.

https://doi.org/10.48550/arXiv.1412.6980