SCIREA Journal of Sociology

http://www.scirea.org/journal/Sociology

May 7, 2023

Volume 7, Issue 3, June 2023

https://doi.org/10.54647/sociology841046

Feature Mining of News Communication Topic Elements Based on BERT Model

Fei Yang Zheng

The University of Chicago, Edward H. Levi Hall 5801 S. Ellis Ave. Chicago, IL 60637, USA

megaedu201503@126.com

Abstract

In order to solve the problems of lack of standardization, fuzzy semantics and sparse features in news topic texts, a feature mining of news communication topic elements based on the BERT model is proposed. In the research, multi-layer fully connected layer feature extraction is performed on the output of the news topic text in the BERT model, and the final extracted text features are purified by the feature projection method to enhance the classification effect. Then the feature projection network is fused in the hidden layer inside the BERT model for feature projection, so as to enhance and purify the classification features through the feature projection of the hidden layer. Experiments are performed on Toutiao, Sohu News, THUC News-L, and THUC News-S datasets. The experimental results show that compared with the baseline BERT method, the two methods have better performance in terms of accuracy and macro-average F1 value, and the highest accuracy is 86. 96%, 86. 17%, 94. 40% and 93.73%, respectively, which verifies the feasibility and effectiveness of the proposed method. It is concluded that the proposed method for news topic text classification combining BERT and FP net is effective and efficient.

Keywords: pre-trained language model; text classification; news topics; BERT; feature projection network

1. Introduction

With the rapid development of information and the emergence of various social media, a large number of complicated information materials are generated [1]. These information and social media are full of different kinds of news. Due to the large number of network users, the news spreads too fast. Once an unexpected social incident occurs, the spread of public opinion will be very fast. If the incident is negative, It will cause huge social public opinions and bring negative impacts [2]. These public opinions are mainly based on news texts and are widely spread on the Internet. Therefore, the classification of news texts is particularly important, and it is the basis for relevant departments to supervise information dissemination. Efficient and accurate classification and recognition of news texts can allow the supervision department to pay attention to the development trend of an event in time. Once the reporting frequency of a certain news is abnormal, it will remind the supervision department to deal with hot events in time to avoid the negative impact of the event fermentation on society [3].

Digital text on the Internet is increasing day by day, from applications to websites, with millions of data, there is a large amount of text data that is difficult for computers to distinguish. Due to the large amount of data and the complex semantics of text, text classification has become a difficult problem. Therefore, how to make a computer classify a large amount of text data is becoming a topic of interest to researchers. Generally, text classification tasks have very few categories. When the classification task has a large number of categories, the traditional Recurrent Neural Network (RNN) (such as LSTM and GRU) algorithms perform poorly in accuracy, so the transformer-based Bidirectional Encoder Representations from Transformers (BERT) model released by Google Brain is used to classify Chinese news texts in the research[4].

At present, the BERT model has become a basic tool. After the further optimization and transformation, it can be widely used in various text mining application scenarios [5]. Since its release, it has attracted the attention of a large number of researchers, and further developed on the basis of BERT, forming a series of models based on BERT optimization and improvement. In the research, various optimization and improvement methods of the BERT model are summarized through the investigation and analysis. By sorting out 41 models, the

following 4 BERT-based optimization and improvement directions and technical routes are proposed. First, a large number of researchers have improved the learning ability of the model for text features by improving the two pre-trained objectives of BERT [6]. Secondly, aiming at the explicit knowledge of a specific domain, a method of integrating external knowledge in the pre-trained model is proposed, which further enriches the text features learned by the model.

2. Literature Review

News text classification includes topic classification and content classification. In the task of news topic text classification, news topic text is usually composed of some words that highly generalize the news content. Due to the lack of standardization and vague semantics, the existing text classification methods perform poorly [7]. The length of news topic texts is short, and it is extremely challenging to extract the complete semantic features of news topic texts of limited length for classification.

News topic classification belongs to the natural language processing (NLP) short text classification task. For the text classification task, it first needs to perform text processing on the relevant text and carry out the text vectorized representation[8]. With the rise of deep learning methods, there are two commonly used word embedding methods, one is the static language model Word2Vec, GloVe; the other is the dynamic language model such as pre-trained model BERT (Bidirectional Encoder Representations from Transformers) and XLNet [9]. The Word2Vec method can better reflect contextual information and is widely used in natural language tasks. The emergence of the pre-trained model BERT solves the polysemy problem that static word vectors cannot solve, and performs well in multiple NLP tasks. News topic classification refers to the feature processing, model training, and output classification of news topics through NLP technology. News topic classification is one of the important research directions of current NLP text classification. Since the development of the Internet, a large amount of news is generated every day, and various news categories are mixed in it. How to better classify it has important research significance.

The vectorized representation of text is to use numerical vectors to represent the semantics of the text, vectorize the text, build a suitable text representation model, and let the machine understand the text, which is one of the core issues of text classification [10]. The naive Bayes model in traditional machine learning does not need to vectorize the text, it records the

conditional probability value of the word, and calculates the conditional probability value of each input word to obtain the predicted value. However, most of the current linear classification models still need to vectorize the text, and a numerical vector must be input to calculate the predicted value. In the traditional feature representation, the bag of words is used to represent the text, which can easily lead to high-dimensional and sparse features, which not only affects the efficiency and performance of text analysis, but also has poor interpretability.

With the development of deep learning, some excellent neural network language models have been proposed, which has greatly promoted the development of the NLP field. In the research, combining BERT and feature projection network (FP net), a news topic classification method BERT-FP net is proposed, which extracts common features through gradient reversal network, and uses feature projection method to extract features from BERT model for feature projection purification. The strong classification features are extracted and the classification effect of news topic texts is improved[11].

3. Methods

3.1 Related technologies

3.1.1 Pre-trained language model BERT

For the BERT model, the bidirectional Transformer encoder is used to obtain the feature representation of the text. The model structure is shown in Figure 1. The training text is input into the multi-layer bidirectional Transformer encoder at the character level for training, and the character-level features of the text are output.

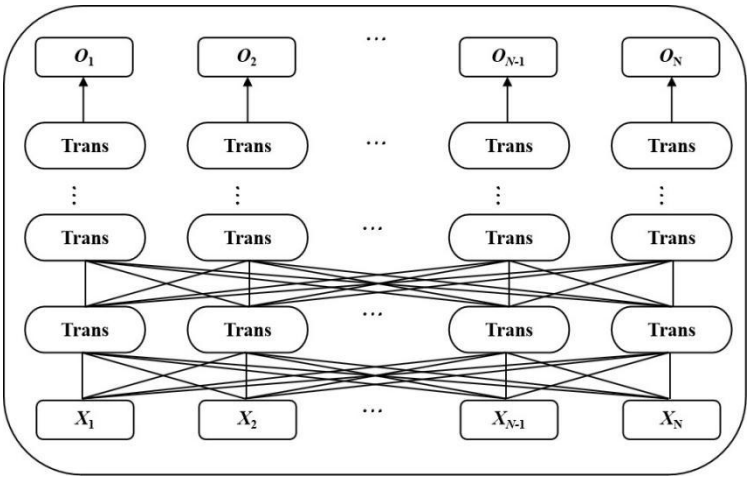

Figure 1 BERT model structure

In the pre-trained stage, the BERT model combines the global visibility of the attention mechanism of the Transformer architecture through the MLM task, which increases the information acquisition of the BERT model, and the random mask makes the BERT model unable to obtain full information and avoids overfitting. The NSP (Next Sentence Prediction) task allows the model to better understand the connections between sentences, so that the pre-trained model is better adapted to downstream tasks [12]. Therefore, the BERT model has a strong ability to understand text semantics and is effective in text classification tasks.

3.1.2 FP net

FP net is a neural network structure that enhances the effect of text classification. It is mainly implemented by using the gradient reversal network, which uses the Gradient Reversal Layer (GRL) to extract the common features of multiple categories [13]. The implementation principle of GRL is introduced in detail and it is used to extract common features in Domain Adaptation. It embeds domain adaptation into the process of learning representations so that the final classification decision can still extract features that are invariant to changes in the domain. FPnet exploits this feature of GRL to extract common features, and adopts a similar adversarial learning approach to improve representation learning through feature projection.

As shown in Figure 2, FP net consists of two sub-networks: the Common feature learning network (C-net) on the right and the Projection network (P-net) on the left.

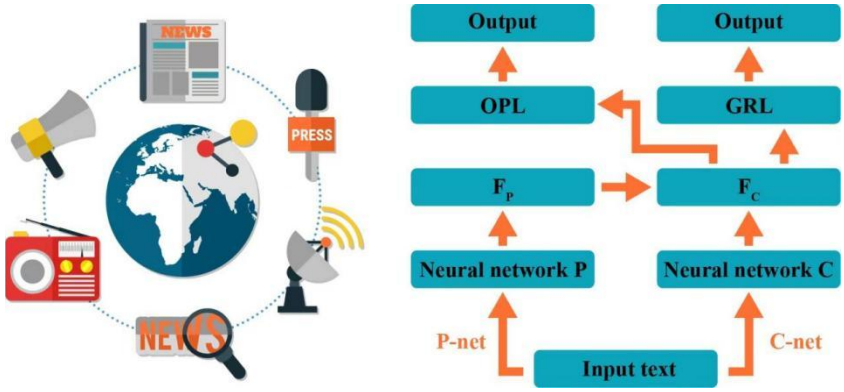

Figure 2 The structure of the feature projection network

The main focus of FP net is to use dual networks for different tasks. The features extracted by the two neural networks are different. Through feature projection, the classification features of the main network are strengthened, thereby improving the text classification effect [14]. FP net can be integrated with existing LSTM, CNN, Transformer, and BERT neural networks. When combining with different neural networks, it is only necessary to replace the neural network P and neural network C feature extractors in the FP net structure with LSTM. , CNN,

Transformer, BERT.As a neural network structure, FP net does not have a fixed form, and its main idea is to strengthen and purify features, so as to achieve the classification effect of strengthening the neural network [15]. Two Text CNN networks are used in Text CNN-FP net as C-net and P-net feature extractors of FP net to extract common features and characteristic features. OPL (Original Projection Layer) is placed after the convolution pooling layer, and feature projection is performed on the last layer of the neural network, thereby improving the classification performance of the Text CNN model.

The addition of GRL to the normal text classification neural network structure in the C-net module will make the feature Fc extracted by the neural network C a common feature [16]. Since the output of C-net is calculated by the loss function, it is affected by GRL inversion during the backpropagation process, so that the loss value of the entire network loss function gradually increases and cannot be classified correctly. The feature Fc extracted by the neural network C is updated in the neural network parameters. In the process, the category information is gradually discarded, and only common information is present, which shows that there is no correct category orientation in the vector space.

In the P-net module, the neural network P performs normal text classification neural network feature extraction, extracts the original text feature F_p, and uses the original feature projection layer OPL to make the original feature F_p and the common feature F_c perform orthogonal projection calculation to obtain more pure classification features F'_p. The feature of F'_p is more clear to the category in the vector space, which can improve the accuracy of the classification task.

The main idea of the OPL layer is vector space orthogonal projection, as shown in Figure 3, F'_p is obtained through F_p and F_c projection, the classification features are purified, thereby enhancing the classification effect.

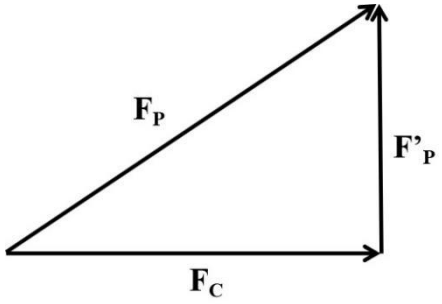

Figure 3 Feature projection

FP net uses dual network cooperation for text classification tasks. The neural network P-net and neural network C-net have the same structure but do not share parameters. The gradient reversal layer GRL is added to C-net, and the feature projection layer OPL is added to P-net. The two networks use the same cross entropy loss function, and the gradient reversal in C-net makes the features extracted by the network not correctly classified, that is, common features are extracted.

Difficulties in the task of news topic text classification mainly include two aspects: 1) The length of the topic text is too short and the semantic information is small, and it is difficult for ordinary text classification models to extract its effective classification semantic information. Some topic words may belong to multiple categories, while other topics may belong to multiple categories. Words cannot point to any category, and it is more suitable to use the BERT model as a feature extractor; 2) Some news contains multiple categories of information, such as financial news and real estate news are usually difficult to distinguish, and technology news is easily confused with automotive news [17]. After using FP net, by calculating the purified vector features, the learned information vector of the input news topic text can be projected into a more discriminative semantic space to eliminate the influence of common features.

Different from the general text classification model, the BERT model can not only use the features finally extracted by the classifier for feature projection fusion, but also can be improved by fused FP net in the hidden layer of the BERT network.

3.2 BERT-FP net framework and its implementation

The BERT-FP net news topic text classification method in the research mainly includes two ways of implementations.

1) BERT-FP net-1. The MLP layer output of BERT-FP net is used for feature projection combination. When using the pre-trained model BERT to build a text classification model, it needs to add an MLP (MultiLayer Perceptron) layer after the BERT output for further feature extraction. The MLP layer uses multiple fully connected networks. [18].

2) BERT-FP net-2. The hidden layers of BERT in the BERT-FP net model are used for feature projection.

The overall model structure of BERT-FP net-1 is shown in Figure 4. The model network is mainly divided into two parts, the left is the BERT projection network P-net, and the right is the BERT common feature learning network C-net.

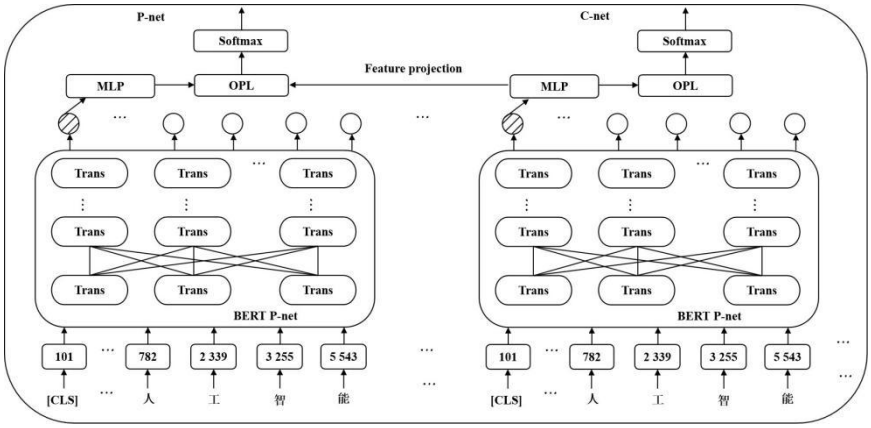

Figure 4 BERT-FP net model framework

The workflow of the BERT-FP net-1 model is as follows. Feature processing is required before the news text is input to the BERT layer, the [CLS] character is added to the beginning of the input news text. And all characters are converted into the corresponding id in the dictionary according to the BERT dictionary, which is input to the BERT model [19]. As shown in Formula (1) and Formula (2).

$$\text{token} = [\text{CLS}] \tag{1}$$

$$ids = [101,782,2339,3255,5543] \tag{2}$$

Since the $[\text{CLS}]$ position vector of the last layer output by the BERT model has global semantic information, here the news text is passed through the BERT model and the corresponding output feature E_{CLS} of $[\text{CLS}]$ is taken out and placed in the MLP layer for further feature extraction to obtain text features E_p and E_c are shown in the following formulas.

$$E_{\text{CIS}} = \text{BERT}(\text{ids})[-1][0] \tag{3}$$

$$E_p = \text{BERT}_p(ids) \tag{4}$$

$$E_c = \text{BERT}_c(ids) \tag{5}$$

The MLP layer contains 2 fully connected layers and an activation function tanh. The dimension parameter of the first fully connected layer is set to [768,768], and the output features are passed through the activation function tanh to enter the second fully connected layer. The dimension parameter is set to [768,class_dim] , class_dim is set according to the number of news text label categories, as shown in the following formulas.

$$f_1 = \tanh\left(f \cdot W_1 + b_1\right) \tag{6}$$

$$f_2 = f_2 \cdot W_2 + b_2 \tag{7}$$

The original features and common features are extracted through the MLP layers of the P-net module and C-net module, respectively, as shown in Formula (8) and Formula (9).

$$f_p = \mathrm{MLP}_p\left(E_p\right) \tag{8}$$

$$f_c = \mathrm{MLP}_c\left(E_c\right) \tag{9}$$

As mentioned above, the C-net module mainly extracts common features. Common features refer to features that do not distinguish between classification tasks. They are common features of all categories. After C-net passes through the MLP layer, the feature extraction is completed, and the features are put into Gradient inversion in GRL. As shown in Formula (10) and Formula (11).

$$\bar{f}_c = \mathrm{GRL}\left(f_c\right) \tag{10}$$

$$\frac{\partial \mathrm{GRL}(x)}{\partial x} = -\lambda I \tag{11}$$

In Formula (10) and Formula (11), λ value is the GRL gradient inversion hyperparameter. The gradient reversal layer does not modify the feature f_c during forward propagation, and passes $-\lambda$ during back propagation to negate the loss function LOSS of the entire C-net network.

The feature projection method is to project the feature vector onto the common feature vector,

and the projection formula is shown in Formula (12).

$$P(e_1, e_2) = \frac{e_1 \cdot e_2}{|e_2|} \cdot \frac{e_2}{|e_2|} \tag{12}$$

Through the feature projection formula, the extracted features are re-projected on the common features. The first projection is to project the original features on the common features, so that the f_p^* vector only contains common semantic information. The second projection obtains the purified feature vector, which only contains classification semantic information, as shown in in Formula (13) and Formula (14).

$$f_p^* = P(f_p, f_c) \tag{13}$$

$$\overline{f_p} = P(f_p, (f_p - f_p^*)) \tag{14}$$

The two networks are identical in structure and do not share parameters. After adding the GRL gradient inversion layer to the C-net, the output of the P-net and C-net both use the Softmax normalized activation function, as shown in in Formula (15) and Formula (16).

$$Y_P = \text{Softmax}(\overline{f_P}) \tag{15}$$

$$Y_c = \text{Softmax}(\overline{f_p}) \tag{16}$$

The dual network is calculated using the cross-entropy loss function. C-net increases the network loss through GRL, and the extracted features cannot be correctly classified, that is, common features are extracted. As shown in Formula (17) and Formula (18).

$$\text{Loss}_p = \text{CrossEntropy}(Y_{\text{truth}}, Y_P) \tag{17}$$

$$\text{Loss}_c = \text{CrossEntropy}(Y_{\text{truth}}, Y_C) \tag{18}$$

In the process of backpropagation, P-net network parameters and C-net network parameters are not shared, Loss $_c$ backpropagation only updates the right C-net network parameters, and Loss $_p$ backpropagation only updates the left P-net network parameters. Although

Softmax and cross quotient loss function are also used in C-net, because C-net performs Loss calculation and back-propagation during back-propagation, it is only for the neural network to obtain common features. The value is the final predicted output of the entire feature projection network.

The main difference of BERT-FP net-2 is that the OPL feature projection layer is between hidden layers inside BERT [20]. The BERT-BASE Chinese pre-trained model is a 12-layer Transformer structure. Since the semantic information extracted from each hidden layer of the BERT model is different, the feature semantic information at the phrase level, syntax level and deep semantic level is extracted from the low level to the high level. While the long-term dependence of text features requires modeling the output of multiple layers of the model. Therefore, in the research, the feature projections of the low, medium and high hidden layers of BERT are combined. And the second feature projection method of the BERT-FP net model, BERT-FP net-2, is proposed through experimental comparison.

BERT hidden layer feature projection is to project the output of the current hidden layer into the next hidden layer. The hidden layer of the BERT-BASE Chinese pre-trained model is 12 layers, as shown in Figure 5. Taking the sixth layer feature projection of the BERT model as an example, the OPL layer is added to the sixth layer of BERT-FP net-2 for feature projection purification, and the BERT-Cnet network structure remains unchanged.

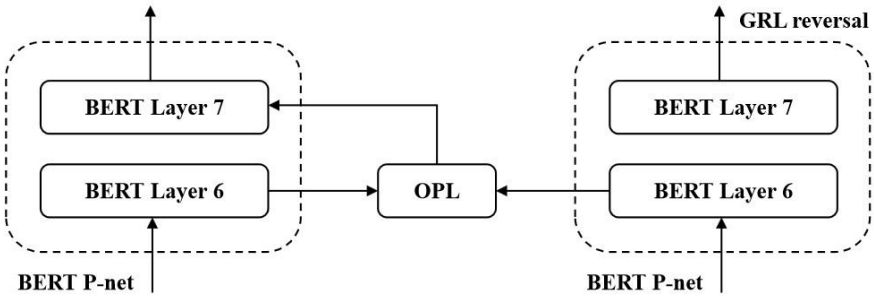

Figure 5 BERT-FP net-2 hidden layer feature projection

Since the BERT model has multiple hidden layers, different hidden layers selected for feature projection experiment comparison through various experiments in the research, so as to obtain the optimal experimental effect.

3.3 Experiment environment and data

The experiment environment of the research is shown in Table 1. In order to evaluate the effectiveness of this model method on the news topic text classification task, four news topic datasets are used to conduct model experiments in the research, as shown in Table 2.

Table 1 Experiment environment

Name	Value
CPU	Intel Xeon Gold 5218
GPU	NVIDIA GeForce RTX5000-16G
Development language	Python-3. 6
Deep learning framework	Pytorch-1.2.0
Development tools	Pycharm-2020.1.3

Table 2 Dataset details

Datasets	Category	Average length	Number of samples			
			Total	Training set	Validation set	Test set
Toutiao	15	22	382688	267878	57409	57401
Sohu News	12	17	34218	22699	5755	5764
THUCNews-L	10	19	200000	180000	10000	10000
THUCNews-S	6	18	60000	48000	6000	6000

1) Toutiao dataset: Toutiao dataset is collected from Toutiao client, including livelihood, culture, entertainment, sports, finance, real estate, automobile, education, science and technology, military, tourism, international, securities, agriculture, e-sports, a total of 15 categories.

2) Sohu News Dataset: The open source Sohu News Dataset is used for data cleaning to remove part of the missing label data in the data, remove the news content, and keep only the news topic. The dataset contains a total of 12 categories including entertainment, finance, real estate, tourism, technology, sports, health, education, automobiles, news, culture, and women.

3) THUCNews-L dataset: THUCNews is generated by filtering the historical data of the Sina News RSS subscription channel from 2005 to 2011, including about 740,000 news documents. In the research, data cleaning is performed on the original data set, and reintegration is divided into finance, real estate, stock, education, technology, society, current affairs, sports, games, entertainment, a total of 10 categories, and each category has about 20,000 pieces of

data.

4) THUCNews-S dataset: THUCNews-S dataset is a small dataset based on THUCNews for data cleaning, including 6 categories of finance, stock, technology, society, current affairs, and entertainment, with 10,000 pieces of data for each category.

3.4 Comparison experiment setup

In order to verify the effectiveness of the news topic classification method combining BERT and feature projection network proposed in the research, 8 classification models with better performance in news text classification are selected for comparison. Among them, TextCNN, FastText, Transformer and DPCNN are combined with Word2Vec word granularity word vector for text classification experiments. ALBERT-FC, BERT-FC, BERT-CNN and BERTBIGRU are combined with pre-trained models for text classification experiments.Details as follows:

1) TextCNN: Multi-window hyperparameters are set to [2, 3, 4], 4 windows can well extract the four-character idiom semantics of Chinese news data, and the number of convolution kernels is set to 256.

2) FastText: The sequence of input text is projected into the word embedding space, and then the text feature vector classification is obtained through the pooling layer. FastText has no convolution operation, and the model structure is simple and fast.

3) Transformer: The encoder is used as the feature extractor. This experiment uses a single-group attention mechanism and 3 encoder blocks as the model composition.

4) Deep Pyramid Convolutional Neural Network (DPCNN): This model refers to the deep residual network (Residual Network, ResNet) to solve the gradient disappearance problem of the deep model. By fixing the number of feature maps, a max pooling operation with a stride of 2 is used to halve the data size of each convolutional layer, and the corresponding computation time is halved, thus forming a pyramid.

5) ALBERT: Using the ALBERT-BASE Chinese pre-training model, the output of the pooling layer in the last layer of the model is connected to the Fully Connected layer (FC) for Softmax classification.

6) BERT-FC: The final [CLS] vector of the BERT model is used to concatenate the FC for classification.

7) BERT-CNN: Each word vector feature output by the encoder of the last layer of the BERT

model is used, and features are further extracted through convolution pooling for classification tasks, where CNN also uses [2, 3, 4] window convolution pooling , the number of convolution kernels is 256.

8) BERT-BIGRU (BERT-Bidirectional Gated Recurrent Unit): The encoder output of the last layer of the BERT model is used to extract the features of each word vector, and the Bidirectional Gated Recurrent Unit (BiGRU) is input to extract the contextual semantic features, so as to perform text classification.

Before the experiment, the four news datasets are preprocessed, non-ASCII characters are filtered out, punctuation marks such as line breaks are cleaned, English characters are case-converted, and Chinese characters are converted between simplified and traditional fonts.

In the comparison experiment, TextCNN, FastText, Transformer and DPCNN models are combined with Word2Vec word granularity word vector to conduct text classification experiments, and respectively train Word2Vec word vector on the training set. In this comparison experiment, the Word2Vec dictionary size is set to 5000, and the character map is 300-dimensional word vector.

In the comparison experiment, ALBERT-FC, BERT-FC, BERT-CNN and BERT-BIGRU are combined with pre-trained models to conduct text classification experiments. ALBERT-FC uses the ALBERT-BASE-CHINESE Chinese pre-training model, and BERT-FC, BERT-CNN and BERT-BIGRU use the BERT-BASE CHINESE Chinese pre-training model. The comparison model hyperparameters are all tuned on the news topic text classification dataset.

3.5 Evaluation indicators

the F1 value of the accuracy rate Acc (Accuracy), the precision rate P (Precision) and the recall rate R (Recall) are used to evaluate the model effect in the research. The calculation formula is as follows.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{19}$$

$$P = \frac{TP}{TP + FF} \tag{20}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{22}$$

Among them, for each news topic classification category, TP indicates that the actual positive sample is predicted to be positive, TN indicates that the negative sample is predicted to be negative, FP indicates that the negative sample is predicted to be positive, and FN indicates that the positive sample is predicted to be negative.

Since this experiment task is a multi-category news topic text classification task, the precision rate, recall rate and the macro averaging value M_F1 value are used as the evaluation indicators. The macro averaging calculation method calculates the precision rate, the recall rate and $F1$ value of each category separately, and then calculates the arithmetic mean for all categories, as shown in Formula (23) and Formula (25). The macro averaging is more suitable as an evaluation metric for multi-class classification tasks.

$$M\_P = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{23}$$

$$M\_R = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{24}$$

$$M\_F1 = \frac{2 \times M\_P \times M\_R}{M\_P + M\_R} \tag{25}$$

3.6 Experiment parameters

The basic parameter settings of the two implementations of the news topic classification method combining BERT and FP net proposed in the research are the same, including the BERT model parameters and the comprehensive model training parameter settings. The BERT model uses Google's open source BERT-BASE Chinese pre-training language Model. The main parameters of the model are shown in Table 3.

Table 3 The main parameters of the BERT model

Name	Value	Name	Value
hidden_size	768	vocab_size	21128
num_attention_heads	12	hidden_act	Gelu
num_hidden_layers	12		

The optimization strategy uses the BertAdam optimizer that is more suitable for the BERT model, the warmup model warmup is set to 0.05, the model learning rate is set to 5E−5, and a

dynamic learning rate strategy is used for learning rate decay with a decay coefficient of 0.9.

Since the average lengths of the four datasets are all around 20, after fine-tuning the length hyperparameters several times, the text input length hyperparameter pad_size=32 is selected, and the gradient reversal GRL hyperparameter λ is set to [0.05, 0.1, 0.2, 0.4, 0.8, 1.0]. As the model training gradient descent changes, common features can be effectively extracted, as shown in Table 4.

Table 4 BERT-FP net model hyperparameters

Name	Value	Name	Value
optimizer	BertAdam	batchsize	128
warmup	0.1	λ	[0.05,0.1,0.2,0.4,0.8,1.0]
learningrate	5E − 5	Dropout	0.5
pad_size	32		

In BERT-FP net-2, feature projection is performed on each hidden layer of the BERT model, and the classification effects of each hidden layer feature projection are compared.

1) Single-layer projection: Feature projection is performed on the 3rd, 6th, 9th, and 12th hidden layers of the BERT model respectively;

2) Double-layer projection: Feature projection is performed on the 3rd, 6th, 9th, 12th hidden layers and the last MLP layer respectively;

3) All layer projections: Feature projections are performed on all 12 hidden layers of the BERT model.

4. Results and Discussions

As shown in Table 5, the BERT-FP net-2 hidden layer feature projection experiment is performed on the Sohu news dataset, 3, 6, 9, and 12 represent the single-layer hidden layer feature projection layer in BERT; 3-MLP, 6 -MLP, 9-MLP, and 12-MLP respectively represent double-layer feature projection; ALL represents feature projection for all layers; MLP is the last MLP layer of BERT-FP net.

Table 5 Experimental results of BERT-FP net-2 hidden layer feature projection on Sohu News dataset

Feature projection layer		Sohu News	
		Acc	M_F1
Single-layer projection	3	0.8243	0.8232
	6	0.8459	0.8466
	9	0.8431	0.8442
	12	0.8617	0.8627
Double-layer projection	3-MLP	0.8431	0.8442
	6-MLP	0.8617	**0.8627**
	9-MLP	0.8386	0.8384
	12-MLP	0.8525	0.8531
All layer projections	ALL	0.8589	0.8604
Way 1	BERT-FP net-1	0.8525	0.8527

In the single-layer feature projection comparison, it can be seen that the feature projection effect of the 12th hidden layer is the best, with the accuracy and F1 value reaching 0.8617 and 0.8627, respectively. From the comparison experiment of double-layer feature projection, it can be seen that the feature projection effect of 6-MLP and 12-MLP layers is the best, but the effect of double-layer projection is lower than that of single-layer 12th hidden layer projection. However, the feature projection effect using all layers for feature projection classification drops more. Comparing BERT-FP net-1, it can be found that the feature projection effect of using the 12th hidden layer in BERTFP net-2 is the best.

In order to further verify the effect of feature projection of BERT-FP net's 12th hidden layer, comparative experiments were carried out on the THUCNews-S dataset. The experimental results are shown in Table 6. It can be seen that the effect of hidden layer projection classification under THUCNews-S dataset is similar to that of BERT-FP net-1.

Table 6 Comparison of BERT FPnet Feature Projection Results on THUCNews S Dataset

Feature projection layer	THUCNews-S	
	Acc	M_F1
12	0.9362	0.9360
BERT-FPnet-1	0.9373	0.9372

The above experiments are compared with the hierarchical feature projection experiments in some hidden layers of the BERT model, and show that the BERT model fusion feature projection layer is suitable for feature projection in the semantic feature extraction layer.

Multiple model experiments and comparison experiments are carried out on the four data sets. The experimental results are shown in Table 7. Among them, BERT-FP net-1 is the feature projection for the final feature output of the model MLP layer, while BERT-FP net-2 is for the final feature output in the model MLP layer. The 12th hidden layer output by BERT is subjected to feature projection and then put into the MLP layer for classification.

Table 7 Experimental results of each model on different datasets

Word embedding	Model	Toutiao		Sohu News		THUCNews-L		THUCNews-S	
		Acc	M_F1	Acc	M_F1	Acc	M_F1	Acc	M_F1
Word2Vec	TextCNN	0.8321	0.7678	0.8320	0.8333	0.9105	0.9107	0.9008	0.9005
	FastText	0.8393	0.7733	0.8236	0.8236	0.9208	0.9209	0.8982	0.8983
	Transformer	0.7939	0.7337	0.7816	0.7814	0.8973	0.8971	0.8845	0.8843
	DPCNN	0.8168	0.7544	0.7705	0.7698	0.9076	0.9076	0.8983	0.8983
ALBERT	ALBERT-FC	0.8460	0.7829	0.8375	0.8384	0.9260	0.9263	0.9105	0.9102
BERT	BERT-FC	0.8559	0.7912	0.8422	0.8416	0.9325	0.9324	0.9227	0.9228

	BERT-CNN	0.8620	0.7965	0.8473	0.8486	0.9421	0.9421	0.9353	0.9350
	BERT-BIGRU	0.8624	0.7981	0.8459	0.8472	0.9352	0.9352	0.9262	0.9262
	BERT-FPnet-1	0.8696	0.8031	0.8525	0.8527	0.9440	0.9438	**0.9373**	**0.9372**
	BERT-FPnet-2	0.8680	0.8011	0.8617	0.8627	0.9410	0.9423	0.9362	0.9360

It can be seen from Table 6 that the two implementations of the news topic classification method combining BERT and FP net proposed in the research are superior to other text classification models in terms of accuracy and macro averaging F1 value, especially better than BERT-CNN and BERT-BIGRU. In order to analyze the performance of each model more intuitively, the experimental results of M_F1 (macro averaging F1 value) of each model are displayed in the form of a histogram, as shown in Figure 6.

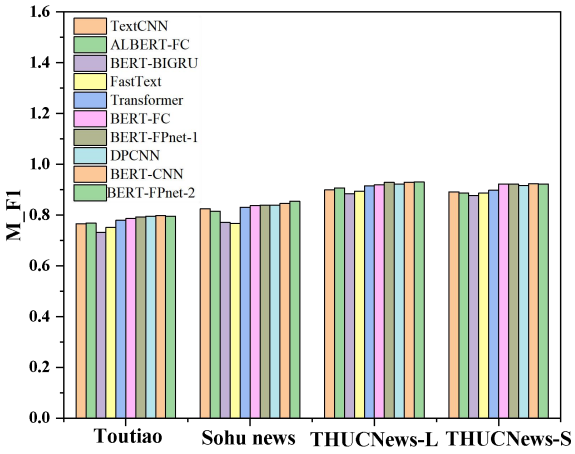

Figure 6 Macro averaging F1 value of each model on different datasets

As can be seen from Figure 6, the effect of the model in the research is better than other comparison models on each dataset, and only the F1 value of the BERT-CNN model is close to the model in the research on the THUCNews-L and THUCNews-S datasets.

And analyzing the data in Table 6, it can be seen that TextCNN, FastText, Transformer, DPCNN using Word2Vec vector, the classification effect is obviously worse than the ALBERT-FC, BERT-FC, BERT-CNN and BERT-BIGRU methods that integrate the pre-training model, indicating that pre-training The language model is better than Word2Vec

in extracting sentence semantic features, which is why the research uses the BERT model to fuse feature projections. Although the ALBERT model innovates on the BERT model and reduces the parameters of the BERT model, it reduces the accuracy of the model to a certain extent.

On the three datasets of Toutiao, THUCNews-L, and THUCNews-S, BERT-FP net-1 has a better projection effect on the MLP layer, while on the Sohu News dataset, BERT-FP net-2 uses the 12th layer of the BERT model. The hidden layer projection effect is better, so different feature projection methods can be selected for different datasets to get the best classification effect.

The parameters of the news topic text classification model proposed in the research that affect the final classification effect mainly include news topic text input length pad_size, GRL gradient reversal parameter λ, and dual network learning rate.

News topic texts generally vary in length, and the model input length pad_size should not be too long or too short. If too short, input length obviously cannot effectively obtain complete semantic information. If too long, the padding value will cause noise, which affects semantic extraction. Due to the characteristics of the attention mechanism of the BERT model, the computational time of the model will also increase exponentially, thus affecting the model classification performance. The main function of the GRL gradient reversal parameter is to help C-net extract effective common features. The dual network learning rate can be divided into synchronous learning rate and asynchronous learning rate during fine-tuning. Synchronous learning rate means that the two networks use the same gradient descent strategy and learning rate, and asynchronous learning rate means that the two networks use different gradient descent strategies and learning rates. Although the dual network optimization strategies of ADam and SGD are used in DANN, and this method is also used in improving the feature projection classification of text. In the research, synchronous learning rate is used to obtain better results.

Parameter comparison experiments are carried out on the THUCNews-S dataset, and the results are shown in Table 8. It can be seen that the value of pad_size varies from the average length of 18 to 40, and the accuracy and F1 value of the model in the research change. It can be seen from the experimental results that when the pad_size value is taken 18, 24, 32 in turn, the accuracy of the model and the F1 value are gradually improved, but when the pad_size value is taken as 40, the model accuracy and F1 value are not effectively improved.

Table 8 Performance comparison of the model in the research on the THUCNews-S dataset under each pad_size

Model	pad_size	THUCNews-S	
		Acc	F1 value
BERT-FP net-1	18	0.9287	0.9278
	24	0.9320	0.9319
	32	0.9373	0.9372
	40	0.9305	0.9305
BERT-FP net-2	18	0.9316	0.9307
	24	0.9357	0.9346
	32	0.9362	0.9360
	40	0.9362	0.9361

The GRL hyperparameter λ takes a static value of 1 and two dynamic λ for experimental comparison. The experimental results are shown in Table 9. It can be seen that different λ values will have a subtle impact on the model classification effect, and a more delicate λ variation range has a better classification effect on the model, and is more helpful for C-net to extract common features.

Table 9 The performance comparison of the model in the research on the THUCNews-S dataset under each λ

Model	λ	Acc	F1 value
BERT-FP net-1	1	0.9343	0.9346
	[0.05,0.1,0.2,0.4,0.8,1.0]	0.9373	0.9372
BERT-FP net-2	1	0.9362	0.9360

FP net-2	[0.05,0.1,0.2,0.4,0.8,1.0]	0.9362	0.9360

In terms of dual network optimization strategy,the different optimization strategies of ADam and SGD used in the feature projection classification of improved text are compared in the research, as well as the dual BERTAdam and synchronous learning rate methods used in the research. The experimental results are shown in Table 10. It can be seen that the method used in the research is better for BERT-based FP net classification.

Table 10 Performance comparison of the model in the research on the THUCNews-S dataset under each optimization strategy

Model	Dual network strategy	Acc	F1 value
BERT-FP net-1	Synchronize	0.9373	0.9372
	Asynchronous	0.9323	0.9323
BERT-FP net-2	Synchronize	0.9362	0.9360
	Asynchronous	0.9353	0.9354

Therefore, in the final experimental comparison part of each dataset, the two gradient descent optimization strategies, Adam and SGD, are used in the dual network structure of FP net with reference to the feature projection classification of the improved text, but the dual BERTAdam optimizer which is more suitable for the BERT model is used. Compared with the baseline BERT method, the above two methods have better performance in terms of accuracy and macro averaging F1 value, with the highest accuracy rates of 86.96%, 86.17%, 94.40% and 93.73%, respectively. The experimental results of mining accuracy are shown in Figure 7.

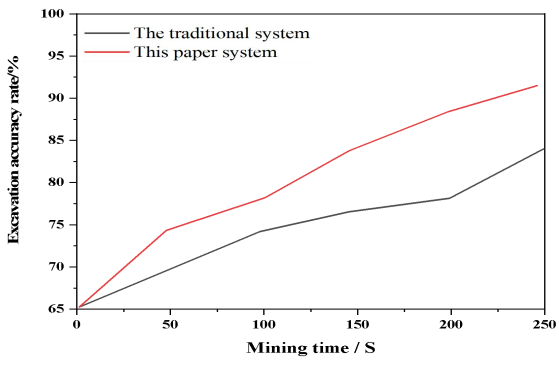

Figure 7 Experimental Results of Mining Accuracy

The system designed in this paper shows that the word meaning of the test news big data dissemination feature keyword is basically the same as that of the real feature keyword, but the expression text of another feature keyword is different; The news big data propagation characteristics and mining system based on data analysis obtained that the keyword matching degree of the propagation characteristics of the test sample is 93%, and a completely incorrect feature appears, and other propagation characteristics are basically the same. The news big data dissemination features based on data analysis and the feature keyword matching degree output by the mining system are 96.5%, and other data dissemination features are basically the same.

5. Conclusions

In the research, the feature mining of news communication topic elements based on the BERT model is proposed. Two news topic text classification methods combining BERT and FP net are proposed. By using the perfect semantic feature extraction ability of the BERT model for news topic texts, the dual BERT model is used to combine the feature projection to complete the news topic text classification task. A GRL gradient reversal layer is added to one of the BERT networks to extract the common features of news topic texts. Then another BERT network OPL is used to project the extracted features on the common features to extract characteristic features and improve the text classification effect. Extensive comparison experiments are conducted on four news topic datasets to verify the effectiveness of the news topic text classification method proposed in the research combining BERT and FPnet.

References

[1] Salminen, J. , Hopf, M. , Chowdhury, S. A. , Jung, S. G. , & Jansen, B. J. . (2020). Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences, 10(1), 1.

[2] Shin, J. . (2020). How do partisans consume news on social media? a comparison of self-reports with digital trace measures among twitter users:. Social Media + Society, 6(4), 173-190.

[3] Hu, W. , Cai, X. , Hou, J. , Yi, S. , & Lin, Z. . (2020). Gtc: guided training of ctc

towards efficient and accurate scene text recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 11005-11012.

[4] Yenduri, G. , Rajakumar, B. R. , Praghash, K. , & Binu, D. . (2021). Heuristic-assisted bert for twitter sentiment analysis. International Journal of Computational Intelligence and Applications, 20(03), 20625-20631.

[5] Liu, Y. , Lu, J. , Yang, J. , & Mao, F. . (2020). Sentiment analysis for e-commerce product reviews by deep learning model of bert-bigru-softmax. Mathematical Biosciences and Engineering, 17(6), 7819-7837.

[6] Sanford, N. , Lavelle, M. , Markiewicz, O. , Reedy, G. , Rafferty, A. M. , & Darzi, A. , et al. (2022). Understanding complex work using an extension of the resilience care model: an ethnographic study. BMC Health Services Research, 22(1), 1-10.

[7] Wang, H. , He, J. , Zhang, X. , & Liu, S. . (2020). A short text classification method based on n-gram and cnn. Chinese Journal of Electronics, 29(2), 248-254.

[8] Levis, M. , Westgate, C. L. , Jiang, G. , Watts, B. V. , & Shiner, B. . (2020). Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychological Medicine, 51(8), 1-10.

[9] Santos, B. , Marcacini, R. M. , & Rezende, S. O. . (2021). Multi-domain aspect extraction using bidirectional encoder representations from transformers. IEEE Access, PP(99), 1-1.

[10] О. А. Розанова, М. В. Кашуба, & М. В. Цинова. (2020). Prospection and retrospection as temporal markers of text organization in the novel "the choice" by n. sparks. Writings in Romance-Germanic Philology(1(44)), 264-272.

[11] Zou, Y. , Shi, Y. , Shi, D. , Wang, Y. , & Tian, Y. . (2020). Adaptation-oriented feature projection for one-shot action recognition. IEEE Transactions on Multimedia, PP(99), 1-1.

[12] Zhou, Y. , Liao, L. , Gao, Y. , Wang, R. , & Huang, H. . (2021). Topicbert: a topic-enhanced neural language model fine-tuned for sentiment classification. IEEE Transactions on Neural Networks and Learning Systems, PP(99), 1-14.

[13] Tang, H. , Mi, Y. , Xue, F. , & Cao, Y. . (2021). Graph domain adversarial transfer network for cross-domain sentiment classification. IEEE Access, PP(99), 1-1.

[14] Wan, X. , Li, Z. , Chen, E. , Zhao, L. , & Xu, K. . (2021). Forest aboveground biomass estimation using multi-features extracted by fitting vertical backscattered power profile of tomographic sar. Remote Sensing, 13(2), 186.

[15] Liu, S. , Whidborne, J. F. , & Chumalee, S. . (2021). Disturbance observer enhanced neural network lpv control for a blended-wing-body large aircraft. IEEE Transactions on Aerospace and Electronic Systems, PP(99), 1-1.

[16] Wen, L. , Li, X. , & Gao, L. . (2020). A new reinforcement learning based learning rate scheduler for convolutional neural network in fault classification. IEEE Transactions on Industrial Electronics, PP(99), 1-1.

[17] Lin, J. P. , Feng, H. S. , Zhai, H. , & Shen, X. . (2021). Cerebral hemodynamic responses to the difficulty level of ambulatory tasks in patients with parkinson's disease: a systematic review and meta-analysis:. Neurorehabilitation and Neural Repair, 35(9), 755-768.

[18] Verma, P. , Awasthi, V. K. , & Sahu, S. K. . (2021). Classification of coronary artery disease using multilayer perceptron neural network. International Journal of Applied Evolutionary Computation, 12(3), 35-43.

[19] Church, K. W. . (2020). Emerging trends: subwords, seriously?. Natural Language Engineering, 26(3), 375-382.

[20] Gtsch, T. , Tuerk, H. , Schmidt, F. P. , Vinke, I. C. , Haart, L. , & Schlgl, R. , et al. (2021). Visualizing the atomic structure between ysz and lsm: an interface stabilized by complexions?. ECS Transactions, 103(1), 1331-1337.