A Study on the EDA-based Classification of News Text

Xu Shuwei¹, Gao Xuyang²*”, Wang Ying³

¹School of Software, Henan University, Institute of Intelligence Networks System, Kaifeng, Henan, China
²School of Software, Henan University, Kaifeng, Henan, China
³School of Software, Henan University, Institute of Intelligence Networks System, Henan Experimental Teaching Demonstration Centre of Modern Network Technology, Kaifeng, Henan, China

*Corresponding author: 104754191399@henu.edu.cn

Abstract—At present, the commonly used word vector methods for news text classification mostly adopt Word2Vec, Glove, Bert and other word vector models, ignoring the remote context connections of Chinese text itself. TextCNN, RNN, BiLSTM and other neural network classification models lack the extraction of important information features of the text, and the word dependence inside the text is not strong, resulting in inaccurate classification results. To solve the above problems, this paper proposes the DPCNN-Attention news text classification model based on ERNIE's pre-training model (EDA for short, the following general). The DPCNN neural network model with Mish() activation function is adopted to obtain the maximum length of semantic association between long distance texts in news texts. By adding attention mechanism into EDA model, in the feature extraction process, according to the importance of words to the classification results, different weights are assigned to them to enhance the word dependence relationship within the text, thus greatly improving the classification accuracy. The EDA model was experimentally verified on the THUCNews dataset. The results showed that the EDA model improved by about 6% compared with BERT's pre-training model, and 0.4% compared with ERNIE's pre-training model. The loss rate decreased by 0.2 compared with BERT and 0.01 compared with ERNIE's.

1. INTRODUCTION

In recent years, with the arrival of the 5G era, the way of information dissemination is becoming more and more rapid and convenient. A variety of portable mobile devices and APP applications continue to emerge. News around the world is more accessible, and more and more people are more willing to obtain various news events of interest through various APP applications and web pages. How to classify a large number of news texts accurately and efficiently and improve users' news reading experience has become one of the important research topics in the field of natural language processing. Based on the existing research results and theories, this paper constructs an EDA model to improve the accuracy of news text classification.

2. MATERIALS AND METHOD

2.1. Related research

The application of text classification in machine learning is mainly reflected in the following four
aspects: (1) Logical regression algorithm. (2) Naive Bayes Algorithm. (3) Random forest algorithms. (4) Support Vector Machine Method. In deep learning, Mnih et al. proposed an attention mechanism approach based on deep learning [1] in 2014 to obtain valid feature information. Experimental results show that the experimental results of RNN-Attention combined network model in network models in image research. Kim proposed TextCNN neural network model [2] in 2015, which proves that the CNN network model has a good effect in the field of natural language processing, but there is a problem that the convolution pool operation cannot obtain the contact information of the context at the same time. Literature the author proposes a cyclic convolution neural network structure RCNN [3], which mainly uses the BiLSTM network to obtain the context information of the input data. The pool layer is mainly used to obtain the key features in the input data, and each exerts its own advantages of different network models. However, only pool layer is retained in convolution layer, which weakens the ability of CNN network model to extract text feature information to some extent. The DPCNN [4] proposed by Tencent AI-lab in ACL2017 is a Word-level and effective deep text classification convolution neural network model.

At present, in order to keep the inner relation of the words in the sentence to the greatest extent, the pre-training model based on large corpus has gradually become a general trend in the field of natural language processing. Google company puts forward the idea of Transformer encoder [5], which can guarantee the long-distance dependence information of input text and combine with the attention mechanism model, which has a great improvement in the computational time complexity. Then the representation (BERT) model [6] of Transformer bidirectional encoder is proposed. BERT model realizes the bidirectional connection of all layers of the model in the real sense. ERNIE model [7] is an improved version of the BERT model. In pre-training, the model parameters are fine-tuned by masking the input words, and the input data are trained by using a large number of Chinese data sets in the pre-training data. Make Chinese text task effect is better.

In recent years, with the development of natural language processing, attention mechanism has attracted more and more researchers' attention. Because the existing network model suitable for NLP has some shortcomings in obtaining text long distance relationship and extracting text key feature information. Therefore, this paper builds a EDA model on the basis of the above contents.

2.2. Model structure
In order to obtain the long distance connection of text more effectively, and to extract the important information in the text effectively, at the same time, to simplify the network structure to a certain extent, the remote and accurately on the basis of simplifying the network structure to a certain extent. Combined with the above content, this paper constructs a EDA model. The EDA model uses the fine-tuning mechanism in the ERNIE model to optimize the model parameters and extract the useful text feature information. Secondly, the ERNIE model results the value is regarded as the input value of the DPCNN layer to further obtain the dependency of the context information of each sentence in the dataset. Finally, the model classifies news text by assigning the corresponding weight to the depth information of the extracted text to capture the local features of the extracted text. The overall structure of the model is shown in Figure 1.
In the aspect of data preprocessing, the text adopts sentence coding. Second, in order to improve the efficiency of the experiment, the length of each piece of data is the same. In this paper, the sentence length is set to 32, the later part of the data length is deleted, the data length is not 32, and the mask mechanism is added to mark the place where each piece of data is 32 and the remaining mark is 0. Finally, the data content and mask value are returned to the contents, used in the following ERNIE pre-training model. The principle of sentence coding is shown in Figure 2:

Mish() activation function is used in both the Attention mechanism and the DPCNN model. In the DPCNN model, the activation function is used to process the data before each convolution operation. Mish() the positive value in the activation function can reach any height, it will not encounter the problem of capping saturation. Theoretically, allowing slightly negative values would have better gradient flow than Relu hard zero boundary. Finally, the smooth activation function allows better information to permeate the neural network, resulting in better accuracy and versatility. The formula is as follows:
This paper uses a Bahdanau Attention, with additive attention mechanism to linearly combine the hidden states of the decoder and all position outputs to obtain context vectors for improving the sequence-to-sequence model. In order to make the obtained information better go deep into the neural network and get better accuracy and generalization, the Mish() activation function is used in the attention mechanism. The attention mechanism layer is used to assign the corresponding weight to the extracted text depth information. Finally, the text feature information of different weights is put into the Softmax function layer for output.

During the Encoder phase, the encoder generates a hidden state vector for each input vector:

\[ h_i = Mish(W[h_{i-1}, x_i]) \]  
\[ o_i = \text{softmax}(Vh_i) \]  

At the Decoder stage, the first step is to generate a semantic vector for that moment:

\[ c_i = \sum_{i=1}^{T} x_i h_i \]  
\[ x_o = \frac{\exp(e_i)}{\sum_{i=1}^{T} \exp(e_i)} \]  
\[ e_i = V^T_i Mish(W[s_{i-1}, h_i]) \]  

Where \( c_i \) is the semantic vector of \( t \) time, and \( e_i \) is the influence degree of the hidden layer state \( h_i \) of encoder at \( t \) time and encoder to the decoder at time and hidden layer state \( s_i \); The \( e_i \) probability is normalized \( x_o \) through the Softmax function. The second step is to pass the hidden information and predict:

\[ s_i = \tanh(W[s_{i-1}, y_{i-1}, c_i]) \]  
\[ o_i = \text{softmax}(V_s_i) \]  

3. RESULT & DISCUSSION

3.1. Experimental environment
Windows10; is the experimental environment GPU NVIDIA Tesla P100 servers were used; CPU Intel (R) Core (TM) i710710 CPU, laptop with 16 G memory. Pycharm Professional; is the compilation platform used Python 3.6, Pytorch version is torch 1.6.0+cu101 and torchvision 0.7.0+cu101.

3.2. Experimental data
The data set used in this paper is a subset of the THUCNews news text classification data set provided by Tsinghua NLP Group, with 200,000 data, the experiment set 180,000 pieces of data to a training set, 10,000 set to check set, ten thousand are set to test sets. Finance, realty, stocks, education, science, society, politics, sports, game, entertainment ten categories. Epoch number set to 5, batch size is set to 128, the sequence length is set to 32, the learning rate is set to 1e-5, the number of hidden layers is set to 768, and the number of dropout is set to 0.5 to prevent over fitting.

3.3. Experimental design and analysis of experimental results

3.3.1. Experimental design
In order to ensure a high accuracy in the paper to EDA model, this paper, under the same condition to run ERNIEBILSTMATT, BERGRU, BERTCNN, ERNIEDPCNN, BERTRNN, BERTDPCNN, BERTRCNN seven different models, from multiple perspectives, the comprehensive comparison, EDA high accuracy of the models.
3.3.2. Experimental Results and Analysis
This paper mainly determines whether the EDA model is better than the results of other models by accuracy, accuracy rate, recall rate, F1 value, loss rate and confusion matrix. The results are as follows:

| Model          | acc    | precision | recall   | f1-score |
|----------------|--------|-----------|----------|----------|
| EDA            | 94.46% | 94.47%    | 94.46%   | 94.46%   |
| ERNIEBILSTMATT | 94.31% | 94.32%    | 94.31%   | 94.31%   |
| ERNIEDPCNN     | 94.07% | 94.09%    | 94.07%   | 94.08%   |
| BERTCNN        | 89.81% | 89.79%    | 89.81%   | 89.79%   |
| BERTRNN        | 89.48% | 89.48%    | 89.48%   | 89.46%   |
| BERTDPCNN      | 88.35% | 88.35%    | 88.35%   | 88.33%   |
| BERTRCNN       | 88.21% | 88.16%    | 88.21%   | 88.17%   |
| BERTGRU        | 88.05% | 88.01%    | 88.05%   | 88.01%   |

EDA model through the experiment on THUCNews news text classification data sets, can be seen from the above a series of charts, EDA model relative to the various models based on BERT pre-training, from the original 88% to 94.46% accuracy, and from the confusion matrix proportion, EDA model relative to other model has improved significantly. In the model based on ERNIE pre-training, the accuracy of the loss function increased from 94.07% to 94.46% when the attention mechanism was added, making a great improvement. Based on the above experimental results, the EDA model proposed in this paper is more accurate than previous models in the classification of news texts, and plays a better role of each module in its corresponding position.

3.3.2.1. Loss function
This paper selects the cross-entropy loss function used in the classification problem. Cross entropy describes the distance between two probability distributions. The smaller the value, the closer the two are. But because the output of the neural network is not necessarily a probability distribution. Hence, Softmax functions are usually used in experiments to transform the forward propagation results of neural networks into probability distributions. Softmax functions are often used in multi-classification processes. The output of multiple neurons is normalized to the interval (0,1), so the output of the Softmax function can be regarded as the probability of classification. The formula is as follows:

\[ C = -\frac{1}{n} \sum_{i} [y \ln a + (1 - y) \ln(1 - a)] \] (9)

\[ a = \sigma(z), \text{where} \quad z = wx + b \] (10)

\[ \frac{\partial C}{\partial \omega_j} = -\frac{1}{n} \sum_{i} (y - \frac{1 - y}{\sigma(z)} \frac{\partial \sigma}{\partial \omega_j}) \] (11)

According to the Sigmoid function \( \sigma(z) = \frac{1}{1 + e^{-z}} \) take the derivative of this \( \sigma'(z) = \sigma(z)(1 - \sigma(z)) \) plug it in to get the final result. This paper evaluates the classification model by judging the size of the loss function. The experimental results are shown in FIG. 3:
The results of Figure 3 show that the loss rate of the EDA model is the lowest with the ERNIEDPCNN model. The EDA model can maintain such a low loss rate on the premise of high accuracy, which is mainly due to the more stable and efficient after adding attention mechanism, DPCNN the model has strong ability to extract local information. From the above results, it can be verified that the EDA model proposed in this paper improves the accuracy and efficiency of news classification.

3.3.2.2. Confusion function

By the experimental results show that the Figure 4 in EDA model, in the late experiment after joined the attention mechanism, through effective feature extraction of the texts and the assigned to the corresponding weights, the EDA model in sorting through effective information, classify the assigned weight has reduced the interference terms the impact of the results of the correct classification is not big, so the results in Figure 4, the confusion matrix of EDA model the main diagonal of the sum of the value is the highest, the same proportion is the highest, so the EDA model classification accuracy is also the highest. The bar chart directly reflects that the EDA model proposed in this paper has higher accuracy in news text classification compared with other models.

4. CONCLUSIONS

This paper proposes a text classification method EDA mixed neural networks, and classifies news texts by using ERNIE pre-training model and neural network combination model DPCNN, attention mechanism. Give full play to the advantages of each neural network model in each position. The experiment shows that the ERNIE pre-training model has a great improvement compared with the experimental results of the BERT pre-training model. After adding the attention mechanism, the accuracy of the classification results has been improved to a certain extent. However, because there are relatively many comparative models selected in this paper, only one data set is selected, and it is not yet separated from the text at a long distance Therefore, the model will be evaluated in different data sets, and at the same time, we should read the literature continuously in reducing the loss rate and
improving the accuracy rate. Get appropriate parameters in the experiment to improve the experimental effect. At present, in the field of natural language processing, the results of most combination models are better than those of single models. During the future NLP development, more attention will be paid to the influence of the combination of multiple models on the experimental results on the basis of the experimental results of the development of a single model.

ACKNOWLEDGMENTS
1. This paper is one of the research results of the project (Project No.:182300410164) supported by the Natural Science Foundation of Henan Province.
2. This paper is one of the research achievements of the Graduate Education Innovation and Quality Improvement Program (No.SYL18020105) of Henan University.
3. This paper is one of the research results funded by the —— Excellence Program (No.SYL19060120) of the Graduate Education Innovation and Quality Improvement Program of Henan University.

REFERENCES
[1] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C] //Advances in Neural Information Processing Systems, 2014: 2204 — 2212.
[2] Kim Y. Convolutional neural networks for sentence classification [EB]. arXiv: 1408. 5882, 2014.
[3] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification [C] //The 29th Conference on Artificial Intelligence (AAAI), 2015: 2267 — 2273.
[4] 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 (Volume 1:Long Papers), 2017: 562-570.
[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st Conference on Neural Information Processing Systems. Long Beach, CA, USA, 2017.
[6] ALEX W, AMANPREET S, JULIAN M, et al. GLUE: a multitask benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018: 353-355.
[7] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [EB/OL] (2018-06-11)[2019-12-23]. https://www.zybuluo.com/songying/note/1307498