Auto-Classification of Similar Categories Based on an Improved BERT-MLDFA Method ——Taking E271 and E712.51 of Chinese Library Classification as an Example

Xiangdong Li  
*School of Information and Management, Wuhan University, Wuhan, 430072, China; Center for E-commerce Research and Development, Wuhan University, Wuhan, 430072, China*

Jian Shi  
*School of Information and Management, Wuhan University, Wuhan, 430072, China*

Qianru Sun  
*School of Information and Management, Wuhan University, Wuhan, 430072, China*

Renxian Zuo  
*School of Information and Management, Wuhan University, Wuhan, 430072, China, renxian@whu.edu.cn*

Follow this and additional works at: [https://aisel.aisnet.org/whiceb2022](https://aisel.aisnet.org/whiceb2022)

**Recommended Citation**

Li, Xiangdong; Shi, Jian; Sun, Qianru; and Zuo, Renxian, "Auto-Classification of Similar Categories Based on an Improved BERT-MLDFA Method ——Taking E271 and E712.51 of Chinese Library Classification as an Example" (2022). *WHICEB 2022 Proceedings*. 66.  
[https://aisel.aisnet.org/whiceb2022/66](https://aisel.aisnet.org/whiceb2022/66)

This material is brought to you by the Wuhan International Conference on e-Business at AIS Electronic Library (AISeL). It has been accepted for inclusion in WHICEB 2022 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Auto-Classification of Similar Categories Based on an Improved BERT-MLDFA Method —— Taking E271 and E712.51 of Chinese Library Classification as an Example

Xiangdong Li1,2, Jian Shi1, Qianru Sun1, Renxian Zuo1*
1School of Information and Management, Wuhan University, Wuhan, 430072, China
2Center for E-commerce Research and Development, Wuhan University, Wuhan, 430072, China

Abstract: The high correlation and low distinction degree in the same low-level category of Chinese Library Classification lead to a high-cost of manual classification. The classification task of such similar texts requires more attention from researchers. Aiming at the difficulty of classifying the highly similar text content in the lower-level data of Chinese Library Classification, this paper proposes a Bidirectional Encoder Representations from Transformers (BERT) with Muti-Layers Dynamic Fusion based on Attention (BERT-MLDFA). The model dynamically integrates the parameters of different levels of BERT through a multi-level attention mechanism, which can capture the subtle semantic information of the text so as to distinguish texts of similar categories better. Its superiority is verified by comparing with baseline models such as Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), BERT, etc. Taking E271 and E712.51 in Chinese Library Classification as the classification targets, our results show that the classification effect of baseline models such as LSTM, CNN and BERT is better than that of traditional machine learning methods such as k-nearest neighbor (KNN), naive bayes (NB) and support vector machine (SVM), among which the put forward BERT-MLDFA model performs the best and the Macro_F1 score reaches up to 0.983. As a result, the put forward model can effectively improve classification efficiency, reduce misclassification, and save large classification task costs.

Keywords: Chinese Library, Classification, Similar content, Deep learning, BERT

1. INTRODUCTION

With the rapid development of the Internet and the proliferation of digital resources, the low efficiency of manual classification in Chinese Library Classification (CLC) has led to the suspension of some digital resource gateways, which urgently needs to be provided effective and efficient automatic classification technology to solve the problem [1]. There are a large number of subdivisions by analogy and imitation in the CLC, that is, the contents of many subordinate categories under the same low-level category are very similar, the semantic correlation degree is large, and the distinction degree is small, which is not only the key and difficult point of manual classification but also the difficulty of the automatic classification task. However, the objects in existing research are mainly in small correlation degree and large distinction degree, which lacks applicability for the classification of similar contents. Therefore, this paper studies the efficiency optimization of automatic classification of similar categories through improving classification accuracy.

In reality, there are a large number of classification tasks with high correlation and low distinction. Some researchers point out that the difficulty of microblog emotional classification is that there are some highly similar but subordinate emotional words. The same word expresses opposite emotions in different semantic environments, which brings great challenges to automatic classification [2]. Actually a key point of emotional analysis is the automatic classification of similar text.
There are also a large number of low-level categories with similar contents in CLC. For example, although the bibliographic information of E271 (Chinese Army) and E712.51 (U.S. Army) is the subject of the army, most of the words used do not explicitly mention the concepts of China or the United States, just can be correctly distinguished through the subtle semantic differences of concepts such as "guerrilla warfare" and "field warfare", "use of grenades in tunnel warfare" and "use of grenades in jungle warfare". The difficulty of similar contents classification leads an urgent need for an efficient automatic classification method to reduce the cost of manual classification in the maintenance of digital resources in CLC.

Therefore, this paper takes E271 (Chinese Army) and E712.51 (U.S. Army), two low-level categories of bibliographic information, as the classification objects of similar categories. Considering the characteristics that BERT can not use all the semantic information learned for automatic classification, an improved BERT with Muti-Layers Dynamic Fusion based on Attention (BERT-MLDFA) model is proposed. The model can better learn the subtle semantic differences between two categories with similar content, and promote the automatic classification between three or more categories with similar content. Therefore, the research on automatic classification of categories with similar content in CLC has strong practical popularization value and research significance.

2. LITERATURE REVIEW

The classification task of CLC is often a multi-classification task. Since the multi-classification problem of three or more categories can be transformed into a binary classification problem through one-to-one decomposition, binary classification becomes the basis of automatic classification. And it has been successfully and efficiently applied to the classification of three or more categories[3]. Therefore, for the research on automatic classification of categories with similar contents, binary classification is the primary solution at present[4].

In most of studies classifying the CLC, the correlation degree between different categories is small, and the distinction degree is large. For example, under the classification system of CLC, traditional machine learning classification algorithms such as k-nearest neighbor (KNN), naive bayes (NB), support vector machine (SVM) are used to automatically classify books, web pages, or other types of documents[5,6]. Some also use deep learning methods, such as Long Short Term Memory (LSTM) is used to classify the literature of University books[7], Convolutional Neural Network(CNN) is used to classify the literature in the national newspaper index[8], BERT is applied on classifying the literature discipline[9]. In summary, the objects of these classifications are generally the books and web pages of the upper-level categories or the subject documents of the middle-level categories. They do not focus on the difficult classification with high semantic similarity in the low-level categories.

The huge classification system of CLC and its unique subdivisions by analogy and imitation mechanism make many subordinate classes with high semantic relevance and low distinction under the same low-level category. The topics between these subordinate classes are very close and difficult to distinguish, causing great difficulties for automatic classification. In terms of classification of similar categories in CLC, there are few relevant studies based on traditional machine learning methods, such as classification of similar categories based on KNN, NB and SVM[10], or combining KNN classification algorithm with the mutual information feature selection method[4]. But a few relevant researches are using the deep learning method.

BERT performs well in classification tasks with similar semantics, but there is still room for optimization. Some microblog emotion analysis studies have obtained the semantic information of the text through deep learning models such as LSTM, CNN and BERT, which has achieved good classification results and BERT performs the best among them[11-13]. BERT with 12 hierarchical structures of bottom, middle and top layers can
use the top structure to learn the semantic feature information of the text well\textsuperscript{[14]}. However, when performing classification tasks, BERT only connects the fully connected layer on the last layer parameters for classification, ignoring the semantic information learned by the bottom and middle layers. To fully use the semantic information learned from all layers of BERT for automatic classification tasks, the BERT-MLF model was put forward to connect the 12 layer structure of BERT through CNN, and has achieved better classification results than BERT\textsuperscript{[15]}. However, the CNN structure in BERT-MLF cannot assign different weights to the semantic information learned by different layers of BERT. When dynamically fusing the parameters of different layers of BERT, the key semantic information may be lost while removing some noise semantic information, resulting in the degradation of classification performance. To fill the gap, this paper assigns different weights to the 12 layer parameters of BERT through multi-level attention mechanism, which can make adaptive weight allocation for key semantic information and noise semantic information, so as to improve the classification effect.

To sum up, the existing research lacks using effective and efficient deep learning model to optimize the classification efficiency of categories with similar contents in the same low-level category of books, documents, or literature. This paper selects two categories with similar contents in the CLC to research the binary classification problem. It puts forward the BERT-MLDFA model that dynamically integrates the parameters of different layers of BERT through the attention mechanism so as to optimize the classification efficiency of text classification task in the low-level categories with similar content.

3. MODEL CONSTRUCTION

3.1 Automatic classification based on LSTM and CNN

The LSTM model has a particular cyclic neural network structure. It uses three gate functions of forget gate, input gate and output gate to obtain the temporal sequences relationship of text. It can obtain context information between text features. CNN is mainly composed of the input layer, convolution layer and pooling layer. The convolution layer gets local information between features through the convolution kernel. LSTM ignores local information between features, and CNN ignores context information between features, so both have advantages and disadvantages in automatic classification. In automatic classification, LSTM and CNN are usually combined with Word2Vec word embedding model to obtain a better classification effect and have become the usual classification method. Therefore, for the automatic classification of similar categories in CLC, this paper first uses the typical deep learning model of LSTM and CNN, combined with the Word2Vec word embedding model, and design a related research framework, as shown in Figure 1.

![Figure 1. Automatic classification framework based on LSTM and CNN](image)

The text classification process based on LSTM and CNN is mainly divided into the following four steps.

First step: Perform word segmentation and stop word removal processing on the text of the training set and test set respectively, and obtain the feature set of each text;

Second step: Use the Word2Vec word embedding model trained on the Wikipedia corpus to perform word embedding representation on the feature set of each text;

Third step: Input the word embedding representation of the training set text into the LSTM and CNN neural network models for training, and obtain a trained neural network model;
Fourth step: Input the word embedding representation of the test set text into the trained LSTM and CNN neural network models for classification prediction, and obtain the classification results.

3.2 Automatic classification based on BERT

BERT is a Bidirectional Encoder Representation from Transformers with multi-layers. A masked language model (MLM) is used to generate a deep two-way language representation. In the pre-training process, the context of the feature is obtained through position encoding. So the dynamic vector representation of the feature is obtained according to the context. It has achieved better results than LSTM and CNN in automatic classification, and has become the current mainstream model. In order to improve the automatic classification effect in low-level categories with similar content in CLC, this paper adopts the BERT model and designs a research framework, as shown in Figure 2.

![Figure 2. Automatic classification framework based on BERT](image)

The automatic classification process based on the BERT model is mainly divided into the following four steps.

First step: Preprocess the training set and test set text according to the input format of the BERT pre-training model to construct the BERT input;

Second step: Initialize the BERT pre-training model weights into the BERT model;

Third step: Input the text of the training set according to the processing result of step one into BERT to obtain the text representation for training, and the model is fine-tuned on the training set to update the weights;

Fourth step: Input the results of the test set text processing according to step into the trained BERT model for classification prediction and obtain the classification result.

3.3 Automatic classification based on BERT-MLDFA

When doing classification tasks, BERT only joins the last layer of parameters and the fully connected layer for classification and ignores the semantic information learned by other layers of BERT. To further improve the classification effect of similar categories in CLC, this paper improves the EBRT model and proposes an improved BERT-MLDFA model. This method fuses the features of different layers of BERT based on the attention mechanism.

Different weights, updated during the training process adaptively, are given to different layer features. So that the semantic information learned by different layers of BERT can be fully utilized, and the feature representation with rich semantic information can be obtained, so that the model can better learn and distinguish texts of low-level categories with similar content. The automatic classification framework based on the improved BERT-MLDFA model is shown in Figure 3.
Features dynamically fused by multi-layers attention in the BERT-MLDFA model is calculated as in Eq. (1) and Eq. (2).

$$t_i = \text{Maxpooling}1D(h_i)$$  \hspace{1cm} (1)

$$C = \sum_{i=0}^{12} \alpha_i t_i$$  \hspace{1cm} (2)

Among them, $\alpha_i$ represents the attention weight of the i-th layer of BERT, $h_i$ represents the hidden state output of the i-th layer of BERT, $t_i$ represents the output of the hidden state of the i-th layer of BERT through Maxpooling1D. C represents fusion multi-layers features based on attention mechanism.

In this paper, for the attention mechanism layer, the input is the 12-layer parameters of BERT after maximum pooling. Different weights are given to the 12-layer parameters based on the attention mechanism to fuse the parameters of different layers of BERT dynamically. Combined with the text Surface features, syntactic features, and semantic features, fusion feature C can capture the subtle semantic differences between texts with high relevance and low discrimination.

The fully connected layer mainly performs full connection and softmax calculations on the features after the attention mechanism layer to obtain the classification probability. In order to prevent the overfitting of the model, this paper adopts twice dropouts, and the probabilities of the two dropouts are respectively 0.1 and 0.2. And the final result takes the average value after two dropouts as the classification probability as in Eq. (3), where p represents the classification probability, W represents the fully connected layer matrix, and b represents the bias.

$$p = \frac{\sum_{i=1}^{2} \text{softmax}(\text{dropout}_i(WC) + b)}{2}$$  \hspace{1cm} (3)

The text classification process based on the BERT-MLDFA model is mainly divided into the following four steps.

First Step: Preprocess the training set and test set text according to the input format of the BERT pre-training model, and construct the BERT input;
Second step: Initialize the BERT pre-training model weights into the BERT model and randomly initialize the attention weights in multi-layers of attention.

Third step: Input the result of processing the training set text according to step one into the BERT-MLDFA model for training, and The model is fine-tuned on the training set to update the weights;

Fourth step: Input the results of the test set text processing according to step one into the trained BERT-MLDFA model for prediction, and obtain the classification result.

4. EXPERIMENTAL DESIGN AND ANALYSIS

4.1 Experimental materials and evaluation methods

The experimental objects of this paper are low-level categories with similar content in CLC. In terms of specific category selection, this paper selects two categories, E271 and E712.51. The two categories are under the E category of the CLC. From the perspective of textual words, most of the words are similar because they are all from army themes. Therefore, E271 and E712.51 are two categories that can better represent many similar sub-categories under the same category that are difficult to distinguish using automatic methods. From the analysis of the architecture in CLC, in terms of secondary categories, the special category classification table of E7 is similar to E2. However, Not exactly the same (this is also the reason why the E2 classification is not directly imitated), but the system is very similar. Specific to the two categories of E271 and E712.51, although it is not a double imitation in the category system, the theme is the army. The categories are only different in regions such as China and the United States. Therefore, they can represent many categories generated by subdivision and imitation mechanisms for regions and themes in CLC. In these two categories, the automatic classification method of the objective test can be effectively applied to the categories generated by other subdivision imitation mechanisms. From the analysis of the rigor of the experiment, it is convenient to compare with the results of published papers that used automatic classification methods and to verify the superiority of the deep learning method compared to the traditional machine learning method in the aspect of classification effect and the effectiveness of the BERT-MLDFA model put forward in this paper. In general, the bibliographic information of E271 and E712.51 are typical experimental objects for the research on automatic classification in low-level categories with similar content in CLC.

To ensure the principle of openness and fairness of the experiment, this paper extracted E271 and E712.51 under the classification number E (military) in CLC from the China Science and Technology Journal Database as the data source of the experiment. Among them, 616 documents in E271 and 1366 documents in E712.51 were collected. Each document includes three parts of information: title, keywords, abstract. And there is no intersection between the two types of text data sets. The text length statistics are shown in Figure 4 (the text length includes the left endpoint and does not include the right endpoint). 80% of the text length is concentrated between 50 and 300.

Figure 4. Data text length statistics
Taking books as an example, even if a large library has 10 million kinds of books (excluding copies), among the more than 50,000 categories of CLC, there are on average less than 200 books in each category. The research on automatic classification methods with few samples must consider the amount of data that can be used in practical applications in the future. To ensure that the experimental results are not affected by randomness and unbalanced data, this paper adopts a balanced data set and divides the experimental materials into five groups. For each group of experimental materials, 200 documents are randomly selected from E271 and E712.51 as the training set. To ensure that the texts in the training set and test set are not duplicated, randomly select 100 documents from the remaining ones as the test set. Experiments were carried out on five groups of experimental materials, respectively. Each group's experimental results were recorded, and the average of the five groups of experimental results was taken as the final experimental result.

To verify the effectiveness of the method proposed in this paper for the classification of categories with similar content, this paper calculates the F1 value based on the combination of precision and recall (Sokolova and Lapalme, 2009). Since the number of texts in the two categories in the experimental materials is equal, the macro-average F1 value (Macro_F1) and the micro-average F1 value (Micro_F1) are consistent, so this paper uses the Macro_F1 value to represent the classification effect of the experiment.

4.2 Experimental environment and hyper-parameter setting

The experiment in this paper is based on the Ubuntu operating system, the size of the video memory is 16G, and the neural network structure is built with the Python programming language and the Torch1.8 deep learning framework. In the preliminary experiments, hyper-parameters based on LSTM, CNN, BERT, BERT-MLF, and BERT-MLDFA models were determined, including learning rate, batch size, training epochs, max sequence length, optimizer, etc. The hyper-parameter setting information is shown in Table 1.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>LSTM,CNN</th>
<th>BERT,BERT-MLF,BERT-MLDFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning rate</td>
<td>1e-3</td>
<td>2e-5</td>
</tr>
<tr>
<td>batch size</td>
<td>60</td>
<td>9</td>
</tr>
<tr>
<td>training epochs</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>max sequence length</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>optimizer</td>
<td>Adam</td>
<td>Adam</td>
</tr>
</tbody>
</table>

4.3 Experimental results and analysis

For the binary classification task in low-level categories with similar content in CLC, this paper sets up two groups of comparative experiments. First, E271 and E712.51 of CLC are automatically classified based on deep learning methods such as LSTM, CNN and BERT, and then study the advantages of deep learning methods compared with traditional machine learning methods such as KNN, NB and SVM; secondly, aiming at the failure of BERT to make full use of the semantic information learned at different layers, to capture the subtle semantic differences between texts with similar content and further improve the classification effect, the parameters of different layers of BERT are dynamically fused based on the attention mechanism. An improved BERT model is put forward called BERT-MLDFA model, compared with the BERT-MLF model based on CNN to fuse the parameters of different layers of BERT, and then analyze the superiority of the improved method put forward in this paper.

The baseline experiments of the first group of comparative experiments are based on the classification effects of traditional machine learning classification algorithms such as KNN, NB, and SVM. Therefore, the best results with different parameter combinations in the literature (Li Xiangdong and Ruan Tao, 2018) are taken as the baseline experimental results compared with the experimental results based on LSTM, CNN and BERT, as shown in Figure 5.
As can be seen from the results in Figure 5, for the binary classification task in low-level-categories with similar content in the CLC, among the traditional machine learning methods of KNN, NB and SVM, SVM performs the best, and the Macro_F1 value is 0.963. Compared with KNN and NB, the Macro_F1 value is respectively increased by 0.4% and 2%; Among the LSTM, CNN and BERT, BERT performs best and the F1 value is 0.980, the Macro_F1 value is increased by 1.4% and 1.6% compared with LSTM and CNN respectively; This paper adopts the three deep learning methods, and deep learning methods are overall better than the three traditional machine learning methods, and the Macro_F1 value of BERT is 1.7% higher than that of SVM.

The second group of comparative experiments compares the classification effects of the BERT, BERT-MLF and the BERT-MLDFA model put forward in this paper. The experimental results are shown in Figure 6.

As can be seen from Figure 6, the BERT-MLDFA model put forward in this paper performs the best with a Macro_F1 value of 0.983. It is 0.3% higher than that of BERT and also 0.2% higher than that of BERT-MLF. BERT-MLDFA further launched a sprint to 1, which is close to 100% correct classification compared with the effect 0.980 of BERT. At the same time, due to the huge number of CLC, for example, for a large library with 10 million kinds of books, in the more than 50,000 categories of CLC, even a 0.3% increase may make 30,000 books classified correctly, which can bring huge time and economic benefits, so it has strong practical
significance.

The following two conclusions can be drawn from the analysis of the above two sets of comparative experiments.

- In the automatic classification of categories with similar content in CLC, the classification effect of the deep learning method is better than that of the traditional machine learning method.
- The improved BERT-MLDFA model put forward in this paper performs the best in the automatic classification low-level-categories with similar content in CLC. When dynamically fusing the parameters of different layers of BERT based on the attention mechanism, it can combine the surface features of the text, Syntactic features and semantic features, which can capture the subtle semantic differences between texts with high correlation and low discrimination. So the superiority of BERT-MLDFA in solving the classification problem in low-level-categories with similar content in CLC is proved.

5.CONCLUSIONS

The classification task in low-level categories with similar content is a significant research direction in the classification system of CLC. Due to the high degree of correlation and a small degree of distinction between texts with similar content, there are only subtle differences in semantic information, which greatly hindered automatic classification. Based on the deep learning methods of LSTM, CNN and BERT models, this paper takes the bibliographic information of the two categories of E271 and E712.51 in CLC as an example to perform binary classifications. The experimental results show that deep learning methods have better classification effects than KNN, NB, SVM traditional machine learning methods. BERT performs better than LSTM and CNN among the deep learning methods, with a Macro_F1 value of 0.980. Since the different layers of BERT learn different semantic information, BERT only uses the parameters of the last layer for automatic classification, ignoring the semantic information learned by other layers, to capture the subtle semantic differences between texts with similar content and further improve the classification effect, this paper proposes an improved BERT-MLDFA model, which dynamically fuses the parameters of different layers of BERT based on the attention mechanism, and can perform adaptive weight allocation for key semantic information and noise semantic information. The experimental results show that BERT-MLDFA The model can obtain subtle semantic differences between texts with similar content, thereby improving the classification effect, and the Macro_F1 value reaches up to 0.983. Further exploration of the application effect of the BERT-MLDFA model in multi-classification with similar content will become the focus of future research.

REFERENCES


