Urban Governance Information Classification Method Based on BERT-TextCNN

Jun Wu[1] , Tianyi Li[1,*] , Xinli Zheng[1]

Corresponding author: 469216198@qq.com ,First author:1161276346@qq.com ,Third author:360114225@qq.com

[1]Hubei University of Technology School of Computer Science, Wuhan Hubei,China

Abstract: In view of the existing Word2Vec traditional word vector model cannot solve the problem of text context semantics, this paper proposes a network model BERT-TextCNN for text classification of urban governance information based on the combination of BERT pre-training model and multi-scale convolution kernel TextCNN convolution neural network. The model obtains text context information more efficiently through rich convolution kernel combination. Through experiments in the urban governance information text data set, the experimental results show that the accuracy rate, recall rate and Micro-F1 value of the model have reached more than 90%, which can effectively improve the classification effect of urban governance information text.

Keywords: BERT,TextCNN, text classification,urban governance information

1. INTRODUCTION

Urban governance data is usually a relevant guidance document, which requires repeated review by staff. It is often inefficient and time-consuming to manually review a specific governance information. At present, there is no process to build structured urban governance data in the field of urban governance. Therefore, it is necessary to study the semantic analysis technology in the field of urban governance data, use the in-depth learning technology to automatically and accurately extract the available urban governance information from the massive urban governance data, help the staff to manage and make decisions, promote the application of administrative intelligence, and prepare for the construction of the urban governance data knowledge map and the urban governance knowledge question and answer system in the future.

In the field of urban governance, Mora Luca et al.[1]explored the interaction model between institutional background, urban digital innovation and urban innovation ecosystem; Dieperink Carel et al[2]; Oleksy Tomasz et al.[3]studied the integrated application scenario of artificial intelligence and urban management.

Research on semantic representation and tag correlation modeling has a wide range of application scenarios, such as information retrieval[4], emotional analysis[5], knowledge question and answer[6], multimodal analysis[7], etc. In the field of multi-classification research, Marilyn Bello et al.[8]used the pooling layer based on two-way association to extract advanced features and labels in the multi-label classification problem; Jianhua Dai et al.[9]defined the label similarity relationship matrix in multi-label space through the label importance weight;

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334281

Wang Jiyao et al.[10]enhance the text representation or reduce the number of labels to optimize the lack of information in the text or the sparsity of the possibility vector.

In recent years, with the rapid development of information technology, the construction of urban governance is gradually deeply integrated with big data and artificial intelligence technology. On the one hand, the application of intelligent digital technology provides effective technical support and guarantee for urban governance, on the other hand, scientific urban governance decision-making also provides a solid policy and theoretical backing for intelligent construction[11]. Urban governance is gradually penetrating into all fields of society, Diversified urban governance models have been derived. For a large amount of urban governance information, multi-classification task is the cornerstone of the whole research process. Therefore, the research content of this paper is of great significance.

2. CONSTRUCTION OF URBAN GOVERNANCE INFORMATION CLASSIFICATION MODEL

2.1 BERT

The structure of BERT model is shown in Figure 1,E1- EN is the text input vector of the BERT model. Next, the semantic features of each word are obtained through the Encoder structure of the 12-layer Transformers. The Encoder structure of the Transformers is represented by Trm. Finally, the output vectors is T1-TN.

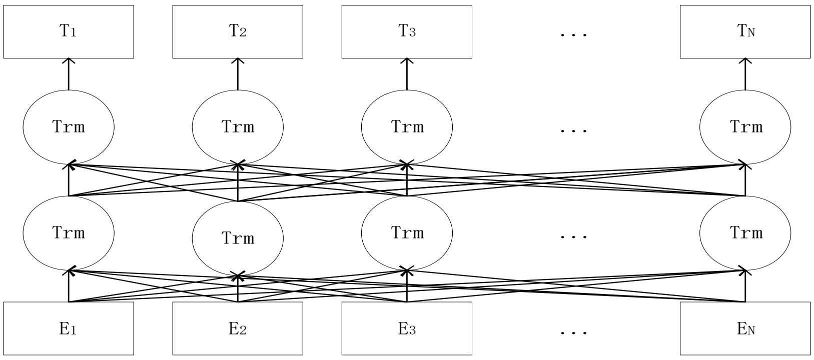

Figure 1. Structure of BERT

2.2 TextCNN

This chapter uses the output of BERT pre-training as the input of TextCNN, and then learns more text features through the convolution layer in TextCNN. The calculation method of feature map H_i is shown in formula (1).

$$H_i = f(W_i * Q_{i:i+k-1} + b_i) \tag{1}$$

Where f is the nonlinear ReLu activation function, W_i is the convolution kernel parameter, Q is the input text matrix, $i:i+k-1$ is the vector spliced by the word embedding vector from position i to position $i+k-1$ in the input sequence, k is the convolution kernel height, and b_i is the offset parameter.

Further, the pooling layer in TextCNN is used to down-sample the feature map, thus reducing the number of parameters and computational complexity of the model. Finally, all pooled feature vectors are connected through the full connection layer, and transferred to the softmax output layer to obtain the classification label with the highest probability. The calculation method of classification result P is shown in formula (2).

$$P = \mathrm{softmax}(W_{\mathrm{p}} * \mathrm{concat}(\mathrm{h}_1, \mathrm{h}_2 ..., \mathrm{h}_n) + \mathrm{b}_p) \qquad (2)$$

Where W_{p} is the weight parameter of the full connection layer, $concat(h_1, h_2 ..., h_n)$ is the connection of all feature maps, and b_p is the offset parameter of the full connection layer.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Data set

Due to the diversified development of urban governance, urban governance documents have certain particularity. The specific governance information in urban governance documents includes one or more types of governance, and the governance types cannot be extracted through governance information. Further semantic analysis of urban governance information is needed. To solve this problem, this paper integrates urban governance data by classification, including the following steps:

(1) Clean urban governance documents and remove invalid information, such as summary statements and thank-you statements.

(2) Sort out the urban governance documents, and integrate them into urban governance information item by item by using the jieba word segmentation package in python through writing procedures.

(3) 12 types of governance are identified, including "economy"、"construction"、"education"、"science and technology"、"livelihood security"、"talent training"、"honor"、"ecological environment"、"social culture"、"problems and deficiencies"、"medical treatment"、"rule of law".

(4) Complete multi-label labeling of urban governance information text and carry out experiments.

3.2 Experimental parameters

The environment for the experiments in this paper is an Inter(R) Core(TM) i5-10400F processor, 8 GB running memory, operating system windows 10, GTX 2060 SUPER GPU, development language Python 3.6, deep learning framework Pytorch 1.5. learning rate 0.00005, random deactivation rate 0.5. Activation function ReLu.

3.3 Evaluating indicator

In this paper, accuracy rate, recall rate and Micro-F1 value are used as evaluation indicators to measure the performance of the model. The Micro-F1 value is an important indicator used to measure the unbalanced test set, taking into account the overall accuracy and recall rate of all labels. The specific formula is as follows:

$$\mathrm{Precision}_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{3}$$

$$\mathrm{Recall}_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{4}$$

$$F1_{micro} = 2 \cdot \frac{\mathrm{Precision}_{micro} \cdot \mathrm{Re}call_{micro}}{\mathrm{Precision}_{micro} + \mathrm{Re}call_{micro}} \tag{5}$$

3.4 Experimental result

In order to study the effect of different sizes of convolution on text feature extraction, this paper first selects convolution cores of different sizes for experiments. The convolution cores of BERT-TextCNN model are set to [3,4,5], [3,4,5,6], [3,4,5,6,7], and the experimental results are shown in Table 1.

Table 1. Comparison experiment with different convolution kernel sizes

Convolution kernel size	$\mathrm{Precision}_{micro}$ (%)	Recall_{micro} (%)	$F1_{micro}$ (%)
[3,4,5]	90.76	90.45	90.61
[3,4,5,6]	90.42	90.12	90.27
[3,4,5,6,7]	90.59	90.28	90.44

As shown in Table 1 above, the effect is best when the convolution kernel size is set to [3,4,5]. Although the difference between the convolution kernel size and the convolution kernel size is [3,4,5,6] and [3,4,5,6,7] is not significant, the increase of the convolution kernel size also increases the calculation cost. Therefore, the convolution kernel size is set to [3,4,5] in this study.

Table 2. Comparative experiment of different models

Model	$\mathrm{Precision}_{micro}$ (%)	Recall_{micro} (%)	$F1_{micro}$ (%)
BERT-TextCNN	90.76	90.45	90.61
BERT-TextCNN-a	90.63	90.23	90.43

BERT-TextCNN-b	90.74	90.45	90.60
BERT-TextCNN-c	90.51	90.25	90.38
WordVec -TextCNN	86.61	87.72	87.16
BERT-Denses	89.96	89.82	89.89
BERT-Seq2Seq	90.36	90.22	90.29

In order to further study the effect of multi-scale convolution kernel combination and the same type of convolution kernel and other mainstream text classification models on text feature extraction in the BERT-TextCNN model, this paper is divided into setting BERT-TextCNN-a model, BERT-TextCNN-b model and BERT-TextCNN-c model, which are respectively set with three types of convolution kernels of the same type, each with a length of 3,4,5. WordVec-TextCNN model uses the traditional Word2Vec word vector representation method to extract text features, and introduces TextCNN model as text classifier. The BERT-Denses model uses the traditional BERT word vector representation method to extract text features, and transforms the multi-label classification problem into multiple binary classification problem prediction. The BERT-Seq2Seq model uses the traditional BERT word vector representation method to extract text features, and provides statement-level feature vectors for the decoder part of the traditional Seq2Seq model. The experimental results are shown in Table 2.

The above results show that the BERT-TextCNN model proposed in this paper has achieved 90% precision, recall rate and Micro-F1 value in the classification of urban governance information, and is superior to other mainstream multi-classification models. In contrast, BERT-TextCNN has shown better classification effect in the classification of urban governance information.

4. CONCLUSION

For the diversified urban governance information, this paper proposes a method to replace the traditional pre-training model with BERT to solve the problem of not being able to effectively extract the context features, and uses TextCNN neural network to construct a classifier, and extracts the text features through multi-scale convolution kernel combination. The experimental results show that good results have been achieved in the classification of urban governance information.

REFERENCES

[1] Mora Luca et al. Smart city governance from an innovation management perspective: Theoretical framing, review of current practices, and future research agenda[J]. Technovation, 2023, 123.

[2] Dieperink Carel et al. City-to-city learning to enhance urban water management: The contribution of the City Blueprint Approach[J]. Cities, 2023, 135.

[3] Oleksy Tomasz et al. What shapes our attitudes towards algorithms in urban governance? The role of perceived friendliness and controllability of the city, and human-algorithm cooperation[J]. Computers in Human Behavior, 2023, 142.

[4] Venkataravana Nayak K. et al. Design of deep convolution feature extraction for multimedia information retrieval[J]. International Journal of Intelligent Unmanned Systems, 2023, 11(1) : 5-19.

[5] Özen İbrahim Akın and Özgül Katlav Eda. Aspect-based sentiment analysis on online customer

reviews: a case study of technology-supported hotels[J]. Journal of Hospitality and Tourism Technology, 2023, 14(2) : 102-120.
[6] Li Wen Zhu et al. Research on Influencing Factors of Knowledge Hiding Behavior in Socialized Q&A Communities: Taking Zhihu as an Example[J]. Complexity, 2022, 2022.
[7] Chowdhary Anuradha et al. Multimodal analysis of the COVID-19-associated mucormycosis outbreak in Delhi, India indicates the convergence of clinical and environmental risk factors.[J]. Mycoses, 2023,
[8] Marilyn Bello et al. Deep neural network to extract high-level features and labels in multi-label classification problems[J]. Neurocomputing, 2020, 413 : 259-270.
[9] Jianhua Dai et al. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation[J]. Knowledge-Based Systems, 2020, 207.
[10] Wang Jiyao et al. Multi-Aspect co-Attentional Collaborative Filtering for extreme multi-label text classification[J]. Knowledge-Based Systems, 2023, 260.
[11] Willis Katharine S. and Nold Christian. Sense and the city: An Emotion Data Framework for smart city governance[J]. Journal of Urban Management, 2022, 11(2) : 142-152.