Research Article

Semantic Analysis of Public Health Medical Issues Based on Convolution Neural Networks

Yulong Xu,1,2 Yingying Zhao,1 Hao Rong,1 Fangfang Liu,1 Yali Lv,1 and Honglei Zhu1

1School of Information Technology, Henan University of Chinese Medicine, Zhengzhou 450046, China
2School of Information Engineering, Zhengzhou University, Zhengzhou 450002, China

Correspondence should be addressed to Honglei Zhu; 26913957@qq.com

Received 12 May 2022; Accepted 4 August 2022; Published 30 August 2022

Academic Editor: Le Sun

Copyright © 2022 Yulong Xu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Text mining and semantic analysis of medical public health issues are the main points for intelligent medical interaction, but less relevant research has been done on them. This article conceives a convolutional neural network for the semantic classification of public health medical issues. The dual convolution layer is used to further reduce the dimension of the data, extract more in-depth information from the data, and map the features. Each convolution layer includes several convolution nuclei to extract semantic characteristics, and then, the complete connection layer is input to the classifier to obtain the results of the classification. To check the classification effect, the dictionary artificial construction and the double hidden-layers neuronal network are used for semantic classification, and the three methods are compared and tested on the six real datasets. The experimental results show that when the quality of the dataset is high, the convolution neural network method proposed in this paper exceeds the last two methods. The proposed method is higher than the construction of the artificial dictionary and the double hidden-layers neural network in the recall rate: 0.153 and 0.037, and greater than 0.07 and 0.01 for the F1 measure rate, respectively. When the quality of the dataset is general, the models of the three methods do not give good classification results. Finally, it is concluded that the convolutional neural network method conceived has a good semantic recognition performance in public health medical issues.

1. Introduction

In recent years, the application of artificial intelligence in the field of medical science has developed rapidly. In smart hospitals such as Internet hospitals and telemedicine, if artificial intelligence technology can be used to intelligently identify and classify the medical and health issues frequently asked by patients during patient consultation and patient telephone consultation, the efficiency and intelligence of the hospital will be greatly improved [1, 2].

Semantic recognition of medical health problems, essentially text semantic analysis, first converts the patient’s question speech into text, then performs the semantic analysis. However, due to the more specialized and complex knowledge involved in the field of medical health, there are fewer studies on semantic analysis of medical health issues [3]. Deep learning is the hottest direction in the field of machine learning in recent years; because of its strong feature extraction ability and learning ability, it has been widely used in image recognition, speech recognition, and natural language processing. Traditional text classification algorithms, such as the k nearest-neighbor (KNN) algorithm [4], decision tree [5], and support vector machine (SVM) [6], have limited fit of complex functions in the case of limited datasets and computational units, which restricts their ability to deal with complex problems, and their learning ability and generalization are not strong [7]. Deep learning transforms the original data into abstract representation layer by layer through multilayer representation learning, automatically learns features from the data, uses its powerful computing and learning capabilities to discover complex structures in high-dimensional data, and the extracted feature information is better used for classification and prediction [8, 9].

Deep neural network models that are widely used in the construction of text classification models include
convolutional neural networks (CNN) [10], recurrent neural networks (RNN) [11], and long short-term memory (LSTM) [12]. As an important network model in deep learning, CNN can classify large-scale text data, and the relevant research results show that CNN has great application value in the field of text classification [13]. For example, the text classification models based on deep convolutional neural networks constructed by Shuai et al. can accurately classify rice knowledge texts with different sample sizes and different complexities [14]. Zhou et al. built a convolutional neural network model to learn and classify medical-related papers based on deep learning text classification methods that can improve the classification accuracy of biomedical texts [15].

Although the above results have achieved good experimental results, few scholars have studied this in the direction of medical public health issue analysis [16–18]. Therefore, to explore the effect and performance of deep learning on public health issues, this paper studies the semantic analysis of medical public health issue texts using improved convolutional neural networks based on the data analysis needs of real medical public health problems.

2. Related Issues and Datasets

Medical health issue semantic recognition, essentially text semantic analysis, analyzes the core meaning of the question according to the text asked by the patient, and then performs subsequent matching answers and processing. The data used in this paper are derived from the Medical Data Mining Algorithm Evaluation Contest sponsored by the Medical Information Branch of the Chinese Medical Association and co-organized by the Medical Big Data and Artificial Intelligence Group of the Medical Information Branch of the Chinese Medical Association.

Specific task: Classification of public health issues. The training set data are shown in Table 1, where question sentence is the patient’s question and consists of a few complete sentences. The question may have six label conclusions, each representing six possible core semantics. The six core semantic labels are as follows: label represents diagnosis; label2 stands for treatment; label3 stands for anatomy/physiology; label4 stands for epidemiology; label5 stands for a healthy lifestyle; and label6 stands for medical choice. If the question sentence contains information about a certain label, the label is set to 1, and if it is not included, it is set to 0, which is a typical multilabel classification problem. Multilabel classification task is that an entity can have multiple labels or be divided into multiple classes at the same time. For example, as shown in Table 1, label 1 and label 2 of the question are set as 1 simultaneously.

In the classification task, a model based on the 4000 training data is built by learning. Then, the model is used to perform semantic analysis on the newly arrived data, and verify indicators such as the accuracy of the analysis.

The training set has a total of 4000 statements, and the frequency statistics of the semantic labels contained in it are shown in Table 2. In Table 2, the frequency of label 1 is 1500, the frequency of label 2 is 2558, the frequency of label 3 is 1, and so on.

3. Design of Convolutional Neural Networks for Semantic Analysis

A convolutional neural network is a multilayer supervised learning neural network that contains an input layer, a hidden layer, and an output layer, which uses the gradient descent method to minimize the loss function to reverse adjust the weight parameters in the network layer by layer and improve the final fit degree through continuous iterative training [19]. The hidden layer generally contains a convolutional layer and a pooling layer, where the function of the convolutional layer is to extract features from the input data, obtain feature mapping, and extract deep features in the data by superimposing multiple convolutional layers. The pooling layer is responsible for compressing features and extracting the main features. The output layer takes advantage of the Softmax classifier. For the problems to be studied, this section uses the dictionary method to represent the question sentence text, and then builds a convolutional neural network model to extract features, learn, and semantically classify the text.

3.1. Data Preprocessing

When training convolutional neural network models to process text, it needs to convert the issue sentence into a vector first. The specific steps: (1) remove the non-Chinese components from the question; (2) look for the question with the longest word count and use its word count as the dimension of the vector, assuming N; and (3) all words that appear in all questions are assigned in the order in which they appear, thus converting the question into a vector dictionary, i.e., key -> value, as shown in Figure 1. After obtaining the dictionary, the words of the question sentence are converted one by one according to the dictionary to obtain a training matrix; if the length of a question sentence is less than N, it is all filled with a value of 0 after the last word, filling the length N. As shown in Figure 2, the first 20 dimensions of data are shown.

The data preprocessing algorithm is given in Algorithm 1.

3.2. Design of Convolutional Neural Networks for Health Issues Semantic Analysis

For the question sentences in test set, they are converted into a matrix numerical vector form according to the dictionary in first, and then and the vector is analyzed by the convolutional neural network. The convolutional neural network model is designed using a double convolutional layer, and each convolutional layer includes multiple convolutional kernels to help extract features.

In the first convolutional layer, multiple convolutional kernels perform feature extraction on the data, where each element that makes up the convolutional kernel corresponds to a weight coefficient and a paranoid amount. Each neuron within the convolutional layer is connected to multiple neurons that are close to the region in the previous layer, and the size of the region depends on the size of the convolutional nucleus. Convolutional kernels regularly scan the input features in their work, multiply the matrix elements in the sensory field, and finally, superimpose the paranois.
After the test set data is processed by the first convolutional layer, it enters the excitation layer.

In the excitation layer, the output result of the convolutional layer is nonlinearly mapped with the Relu() function to improve the fitting degree of the nonlinear case, and after nonlinear processing, the data are fed into the pooling layer.

The pooling layer uses maximum pooling to extract the maximum value of a location and its adjacent areas, then uses the maximum value as the value of the location to reduce the dimensionality without changing the depth of the data.

After the above three layers of processing, the output data are again fed into the second convolutional layer, the excitation layer, and the pooling layer. After two convolutions, the data dimension is further reduced, the number of parameters and weights are reduced, and the deeper information of the data can be extracted to obtain feature mapping. The data are then expanded into one-dimensional data, which is fed into the fully connected layer. The fully connected layer nonlinearly combines the features extracted from the convolutional and pooled layers to output them to the output layer. See the description of Algorithm 2 for details.

The algorithm uses the binary cross-entropy loss function BCEWithLogitsLoss(), which integrates the sigmoid of the output layer with the loss function BCELoss for multilabel classification problems.

4. Manually Constructing the Dictionary Method and Double Hidden-Layers Neural Network

To verify the effectiveness of the proposed method, this section uses manual dictionary construction and neural network methods to solve this problem.

The manual construction dictionary method extracts the words in each category manually, which removes the words with weak medical relevance and saves the words with strong medical relevance. Then, a thesaurus database is established, and the test data can be compared with the lexical database. A dual hidden-layer neural network consists of the input layer, the hidden layer, and the output layer. The input dimension is the dimension of the data, the output

Table 1: Training set data description.

| Field name              | An example of a field                                                                 |
|-------------------------|--------------------------------------------------------------------------------------|
| ID                      | QC0000000012                                                                         |
| Question sentence       | Total cholesterol is 7.38, what level does it belong to? Description of the condition (incidence time, main symptoms, symptom changes, etc.): 1 and 2 years of history, with mild cerebral infarction. Age 67. Thank you. In the service of rosuvastatin, one piece every other day. All right? |
| Label1 (diagnosis)      | 0                                                                                     |
| Label2 (treatment)      | 0                                                                                     |
| Label3 (anatomy/physiology) |                                               |
| Label4 (epidemiology)   | 0                                                                                     |
| Label5 (healthy lifestyle) |                                               |
| Label6 (medical choice) | 0                                                                                     |

Table 2: Semantic label statistics.

| Label  | Frequency |
|--------|-----------|
| 1      | 1500      |
| 2      | 2558      |
| 3      | 1         |
| 4      | 596       |
| 5      | 56        |
| 6      | 46        |

Figure 1: Dataset dictionary.

Figure 2: Training matrix.
dimension is the category number of classification, and the output layer uses the Softmax classifier [14].

4.1. The Manual Construction Dictionary Method. There are 5000 public health issues, 4000 of which are used to build dictionaries, and the remaining 1000 are used to verify the method’s performance.

When constructing a dictionary, establish a dictionary of stop words to process the question sentence first. The dictionary of stop words contains commonly used modal words and medically weak words that may have a negative effect on model performance. So, when we process questions, we remove all the words from the question sentence that are in the dictionary of stop words. Figures 3 and 4 are the results before and after processing, respectively.

Manually constructed dictionary analysis methods mainly include cleaning and parting words for each public health issue in the dataset first. Then, take the thesaurus set of six labels: consider a label, if the label conclusion value is 1, then the corresponded question data after processing is proposed, and all the extracted data are combined and summarized to obtain the thesaurus set of the label, a total of six thesauruses are obtained.

Finally, comparing the classification, the test data to be classified are compared with the six thesauruses, and if one of the words of the data exists in the thesaurus, the corresponding label is set to 1. The algorithm framework is as follows:

4.2. The Double Hidden-Layers Neural Network. The double hidden-layers neural network method designed in this paper contains the following steps. First, the regular matching of all the problems is to remove non-Chinese characters, count all the different words, and build a dictionary. The training data
are then converted into a training matrix based on the
dictionary. Then, build a double hidden-layer neural net-
work model, and train the model to complete the prediction. 
Finally, get the thesaurus set for the six categories:

Input: Public health issues training and test set 
Output: The test set predicts the results 
Begin 
(1) Read the training set and test set 
(2) Set jieba external library stop words and custom word sets 
(3) Part words and clean training and test sets based on stop words and custom dictionaries 
(4) Get the thesaurus set for the six categories 
for six categories: 
define the theology list $X_n$ 
end for 
for all data for a single category: 
if the corresponding category label is 1: 
add the phrases for the row data to list $X_n$ 
end if 
end for 
(5) Clear duplicate words from $X_n$ 
(6) Predict the test set and get performance metrics 
End

Algorithm 3: Manually build a dictionary for semantic analysis.

Description: The patient was a typical “three high”, wanting to take aspirin as a preventive medication, but devel-
oped allergic symptoms. Once treated situation and effect: before high blood pressure, three years ago checked out diabetes. Long-term use of metformin and blood pressure medication. How to GET Help: The patient is typically 
"three high", wanting to take aspirin as a prophylactic but developing allergy symptoms. Can you use Xuesaitong 
dispersible tablets instead of white aspirin as preventive medication for a long time? Are there any other drugs 
recommended? ’

Figure 3: Data before processing.

["Description", "patient", "typical", "three", "high", "take", "aspirin", "preventive", "medication", "developed", "allergic", "symptoms", "treated", "situation", "effect:", "blood", "pressure", "diabetes", "Long-term", "metformin", "prophylactic", "developing", "allergy", "symptoms", "Xuesaitong", "dispersible", "tablets", "drugs", "recommended"]

Figure 4: Data after processing.

for $i = 1, \ldots, N$ do
Step1. Hidden1:
(1) Enter the data with a dimension of 1622
(2) The function Relu() is used for the nonlinear processing of the data 
(3) The output data dimension is 1622 
Step2. Hidden2:
(1) Enter the data obtained by Hidden1 
(2) The function Relu() is used for the nonlinear processing of the data 
(3) The output data dimension is 811 
Step3. Out:
(1) Enter the data obtained by Hidden2 
(2) The function Relu() is used for the nonlinear processing of the data 
(3) The output data dimension is 6 
end for 
Step4. 
Based on the above results, the test data are multilabeled classification using the loss function BCEWithLogitsLoss, obtaining the classification results. 
End

Input: Training dataset $X$ and test dataset 
Output: The results are predicted based on the trained model 
Begin
5. Experimental Comparative Analysis

The experiment was conducted under the Windows 10 system with the processor: Intel(R) Core(TM) i5-8300H CPU @2.30 GHz GPU: NVIDIA GeForce GTX 1050Ti 4G. The memory size is 16 GB. The programming language is Python 3.8.3, the development tool is Jupyter Notebook, the deep learning framework is PyTorch 1.6.0, and the Cuda version is 10.2.

5.1. Text Training. There are a total of 5,000 pieces of data in the public health issue dataset, of which 4,000 pieces of data are used as training sets and the remaining 1,000 pieces of data are used as test sets. Train a double hidden-layer neural network model and a convolutional neural network model on the training set and Figure 5 adjusts the relevant parameters. The method of manually constructing dictionaries is established and learned by humans. Finally, 1000 pieces of test sets are used in Figure 6 to test the performance (Figure 7) of the model. The loss and accuracy of Figure 8 training Figure 9 two types Figure 10 of neural networks on Figure 11 a training set of 4000 pieces of data from public health issues vary with the number of iterations as shown in and the training set contains 4000 pieces of data.

In Figure 5, as the number of iterations increases, the loss value of the double hidden-layer neural network model tends to be 0 and the loss value of the convolutional neural network model tends to be 0.32.

In terms of the changing trend of accuracy, with the increase in the number of trainings in Figure 6, the accuracy of the double hidden-layer neural network tends to be 1 and the accuracy of the convolutional neural network tends to be 0.68. In Figure 7, the accuracy of the hidden-layer neural network tends to be 1 and the accuracy of the convolutional neural network tends to be 0.65. In Figure 8, the accuracy of both types of neural networks rapidly tends to be 1. Figure 9 shows that the accuracy of the double hidden-layer neural network tends to be 1 and the accuracy of the convolutional neural network tends to be 0.88. Figure 10 shows that the accuracy of the double hidden-layer neural network tends to be 1 and the accuracy of the convolutional neural network tends to be 0.91. Figure 11 shows that the accuracy of the double hidden-layers neural network tends to be 1 and the accuracy of the convolutional neural network tends to be 0.84. The two curves in Figure 8 coincide and tend to be 0.96.

According to the above results, for the hidden-layer neural network, it can be seen from Figures 6 to 11 that the accuracy rate tends to be 1. At the same time, combined with the loss curve in Figure 10, the loss of the double hidden-layer neural network tends to be stable slowly, and the loss is close to 0. After analysis, the double hidden-layer neural network structure is relatively simple, and the data can be fit quickly. However, the double hidden-layer may have the risk of overfitting. Regarding convolutional neural networks, the accuracy of Figures 6 and 7 is finally located at 0.65, indicating that the generalization ability of the model still has a large room for improvement on label1-2. The accuracy of Figures 9 to 11 is around 0.9, manifesting that the convolutional neural network can well learn the potential relationship between data on the label4-6. The reason for the phenomenon that the accuracy of the convolutional neural network in Figures 8 to 11 is significantly higher than that of Figures 6 to 7 is that the label4-6 of data as 1 in the dataset is less than the label1-2 of data.

After analysis, the reason for the curve coincidence phenomenon of label3 can be seen: in the training set of 4000 data, only the label3 of one data is 1 and the label3 of the other 3999 data is 0, resulting in the learning of the two types of neural network models too fast, so the accuracy curve coincides.

5.2. Effect Assessment. After the above model and manual dictionary method are trained, the classification performance of the model is tested on the test set. The classification performance of each model is evaluated using precision,
recall, and F1 values. The result is shown in Figures 12–14. (Test set contains 1000 pieces of data.)

In Figure 12, the three models are closer to the precision of label1, near 0.36. Label2 is also closer, at around 0.66. The precision of label3-6 is lower. Overall, the classification effect of the three models on label2 is better.

In Figure 13, the recall of the manually constructed dictionary on label1 reaches 0.56, while alternative models are less effective. The model of the double hidden-layers neural network and the convolutional neural network on label2 is better, in which the convolutional neural network is slightly better than the double hidden-layers neural network. On label3, the indicator is all 0. On label3-6, the effect of two types of neural network is poor, while the method of artificial dictionary construction is stable, and the index is located at about 0.50.

In Figure 14, the F1 score of the manually constructed dictionary on label1 reaches 0.44, the double hidden-layers
neural network reaches 0.42, and CNN reaches 0.26. The index of convolutional neural networks on label2 is up to 0.688. All three models perform poorly on the label456.

An analysis of Figures 12 to 14 shows that only on label2, the performance of the three models is relatively stable, and the performance of convolutional neural networks is slightly better than that of double hidden-layer neural networks. The study found that label2 in the test set had a large amount of 1 data. In label3, there is no 1 data in the test set, resulting in a performance indicator of label3 that is all 0, and the model cannot predict it correctly. The data with a 1 in label1 is less than label2 and more than label3-6, resulting in the phenomenon that label2 has the best classification effect, label1 is weak, and label3-6 is very poor. For the method of artificial dictionary creation, the key lies in the creation and optimization of deactivated dictionaries, which is more limited by the medical literacy of the screener.

6. Conclusion

Experimental results show that in the case of excellent datasets, convolutional neural networks have better performance in the classification of public health issues than double hidden-layer neural networks and manually constructed dictionaries. However, there are still shortcomings in this experiment, such as the preprocessing of text and the establishment of the training matrix, without considering the weight of specific words or words and the relationship between words and words, the stability of each trained model is poor, and the good model that has not yet been trained can achieve a more stable classification effect in label1 to label6. For future experiments, we can study the preprocessing of text and the construction of the training matrix deeply. This experiment helped the author realize the research and application of the convolutional neural network in the field of public health issue classification and medicine, and the convolutional neural network can greatly improve the efficiency and intelligence of hospitals to some extent. And the model obtained by training can assist doctors to make medical judgments and promote the better integration of medicine and computer technology.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.
Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (81703946, 61902113, and 81973791), the subproject of the National Key Research and Development Program (2017YFC1703506), the Science and Technology Research Project of Henan Province (212102310362), the Young Teacher Program of Higher Education Institutions of Henan Province (2020GGJS104), and Ph. D Foundation of Henan University of Chinese Medicine (BSJJ2022-15).

References

[1] X. Kong and R. Wang, "Artificial intelligence and its application in medical field," Journal of Medical Informatics, vol. 37, no. 11, pp. 2–5, 2016.

[2] L. Meng, W. Tan, J. Ma, R. Wang, X. Yin, and Y. Zhang, "Enhancing dynamic ECG heartbeat classification with lightweight transformer model," Artificial Intelligence in medicine, vol. 02, 2022.

[3] B. Liu, E. Blasch, C. Yu, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naive bayes classifier," in Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, October 2013.

[4] R. C. Ambagtsheer, N. Shafabady, E. Dent, C. Seiboth, and J. Beilby, "The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set," International Journal of Medical Informatics, vol. 136, 2020.

[5] A. Kim and H. Kim, "A New Classification Tree Method with Interaction Detection Capability," Computational Statistics & Data Analysis, vol. 165, 2022.

[6] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[7] F. Debole and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets," Journal of the American Society for Information Science and Technology, vol. 56, no. 6, pp. 584–596, 2005.

[8] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[9] W. Tan, L. Pan, X. Li, S. Xu, Y. Chen, and J. Yang, "Segmentation of lung airways based on deep learning methods," IET Image Processing, vol. 01, pp. 1–13, 2022.

[10] H. Cai, Research on Short Text Classification Method Based on Convolutional Neural Network, Southwest University, Chongqing, China, 2016.

[11] Q. Gong, Text Classification Based on Recurrent Neural Network Model, Huazhong University of Science and Technology, Wuhan, China, 2016.

[12] M. Zhao, H. Du, C. Dong, and C. Chen, "Diet health text classification based on word2vec and LSTM," Transactions of the Chinese Society for Agricultural Machinery, vol. 48, no. 10, pp. 202–208, 2017.

[13] Aruna, Research On Text Classification Based on Improved Convolutional Neural Network, Inner Mongolia University for Nationalities, Tongliao, China, 2020.

[14] F. Shuai, T. Xu, Y. Zhou, D. Zhao, N. Zhao, and H. Wang, "Rice knowledge text classification based on deep convolution neural network," Transactions of the Chinese Society for Agricultural Machinery, vol. 52, no. 3, pp. 257–264, 2021.

[15] Y. Zhou, Z. Cui, S. Fan, and X. An, "Deep-learning - based biomedical text classification," Chinese Journal of Medical Library and Information Science, vol. 28, no. 11, pp. 1–10, 2019.

[16] Z. Hu, Z. Zhang, H. Yang, Q. Chen, and D. Zuo, "A deep learning approach for predicting the quality of online health expert question-answering services," Journal of Biomedical Informatics, vol. 71, pp. 241–253, 2017.

[17] Z. Hu, Z. Zhang, H. Yang, Q. Chen, R. Zhu, and D. Zuo, "Predicting the Quality of Online Health Expert Question-Answering Services with Temporal Features in a Deep Learning Framework," Neurocomputing, vol. 275, 2017.

[18] A. Luo, Z. Xin, Y. Yuan et al., "Multi-dimensional feature classification of health information needs of hypertensive patients in the online health community: analysis of 1, 000 patients' question records," Journal of Medical Internet Research, vol. 22, no. 5, 2019.

[19] G. E. Hinton and R. Salakhutdinov, "Replicatedsoftmax: an undirected topic model," Advances in Neural Information Processing Systems, pp. 1607–1614, 2009.