PLOS ONE

RESEARCH ARTICLE

A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications

Xiaoli Li[1], Yuying Zhang[1], Jiangyong Jin[1], Fuqi Sun[1], Na Li[2], Shengbin Liang[1,3]*

1 School of Software, Henan University, Kaifeng, China, **2** School of Digital Arts and Communication, Shandong University of Art & Design, Jinan, China, **3** Institute for Data Engineering and Science, University of Saint Joseph, Macao, China

* liangsbin@henu.edu.cn

Citation: Li X, Zhang Y, Jin J, Sun F, Li N, Liang S (2023) A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications. PLoS ONE 18(3): e0282824. https://doi.org/10.1371/journal.pone.0282824

Editor: Muhammad Fazal Ijaz, Sejong University, REPUBLIC OF KOREA

Received: November 21, 2022

Accepted: February 23, 2023

Published: March 16, 2023

Data Availability Statement: The data used in this manuscript involves third-party data: webMedQA and CMDD, but all these data are public. webMedQA is a real-world Chinese medical question answering dataset collected from online health consultancy websites. The dataset can be downloaded from the GitHub repository (https://github.com/hejunqing/webMedQA). CMDD called Chinese medical dialogue data, the dataset can be downloaded from the GitHub repository (https://github.com/Toyhom/Chinese-medical-dialogue-data).

Abstract

Recently, a lot of Chinese patients consult treatment plans through social networking platforms, but the Chinese medical text contains rich information, including a large number of medical nomenclatures and symptom descriptions. How to build an intelligence model to automatically classify the text information consulted by patients and recommend the correct department for patients is very important. In order to address the problem of insufficient feature extraction from Chinese medical text and low accuracy, this paper proposes a dual channel Chinese medical text classification model. The model extracts feature of Chinese medical text at different granularity, comprehensively and accurately obtains effective feature information, and finally recommends departments for patients according to text classification. One channel of the model focuses on medical nomenclatures, symptoms and other words related to hospital departments, gives different weights, calculates corresponding feature vectors with convolution kernels of different sizes, and then obtains local text representation. The other channel uses the BiGRU network and attention mechanism to obtain text representation, highlighting the important information of the whole sentence, that is, global text representation. Finally, the model uses full connection layer to combine the representation vectors of the two channels, and uses Softmax classifier for classification. The experimental results show that the accuracy, recall and F1-score of the model are improved by 10.65%, 8.94% and 11.62% respectively compared with the baseline models in average, which proves that our model has better performance and robustness.

Introduction

Recently, machine learning and deep learning methods have been widely used in the precision medicine field. For example, El-Sappagh et al. [1] proposed a novel ensemble learning framework for Alzheimer's disease progression incorporating heterogeneous base learners into an integrated model using the stacking technique, Ali [2] et al. proposed a smart healthcare system for heart disease prediction using ensemble deep learning and feature fusion approaches.

Funding: This work was supported in part by the postdoctoral fund of FDCT, Macau (0003/2021/APD) and Key scientific research projects of universities in Henan province, China (23B520005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

F. Ali et al. [3] proposed a new medical monitoring framework based on cloud environment and big data analysis engine to accurately store and analyze medical data and improve classification accuracy. Alfian et al. [4] utilized BLE-based sensor devices and real-time data processing, and uses machine learning algorithms to help diabetic patients better self-manage their chronic diseases. Srinivasu et al. [5] proposed a deep learning-based MobileNet V2 and Long-term Short-Term Memory (LSTM) for classifying skin diseases. Meanwhile, online communities have provided a platform for patients to seek medical treatment and medicine. With the continuous enrichment of medical dialogue text in the social network, analyzing and mining medical text to recommend medical departments for patients is of great significance for improving their experience and diagnosis efficiency.

Medical text is embedded with a large number of medical nomenclature and symptom descriptions. Mining and analyzing medical text can predict diseases and classify patients. Medical text classifications have been applied in various fields of precision medicine, including clinical text classification [6–8], disease features extraction [9, 10], disease features and pathological status [11], phrase detection for medical text [12], extracting biomedical data [13, 14], disease prediction [15, 16], healthcare [17–19], identification of drug interactions [20–23] and identification of adverse drug events [24, 25].

Traditional text classification methods use support vector machine (SVM) [26], Naïve Bayes [27] and other methods to train text word segmentation, statistical word frequency. However, due to the huge dimension of feature vector and the lack of understanding of context, the classification efficiency and accuracy is low. Text classification methods based on deep learning benefit from the powerful capabilities of text feature representation and feature extraction, obviously, these methods are generally better than traditional methods. Yoon Kim [28] proposed TextCNN model, and CNN was first used in the tasks of text classification. On the premise of maintaining the structure of CNN, TextCNN simplifies the number of convolutional layers and input parameters, which effectively improves the training efficiency, but the model has poor interpretability. Liu et al. [29] proposed the TextRNN model, which took the word order and contextual information into account in the feature extraction process. It retained part of the "memory" through the output links and made comparative understanding for long text processing. Lai et al. [30] put forward RCNN model, which combined two-way RNN and maximum pooling operation, made up for the deficiency that one-way RNN only considers the past time information, and paid more attention to the important information of the text through maximum pooling, thus achieving excellent classification results. However, the RNN cycle mechanism is too simple, and it is easy to perform continuous multiplication operations in time when the gradient is backpropagated, resulting in the problems of gradient disappearance and gradient explosion. The improved models such as LSTM and GRU of RNN alleviate the problems of gradient disappearance and gradient explosion to a certain extent. Therefore, LSTM and GRU are used to replace RNN to obtain the contextual information of text [31, 32].

Medical community text often includes topics such as online consultation, fitness and weight loss, intractable diseases, coronavirus disease, and health care. Chinese medical text has a complex structure, contains professional knowledge and has temporal characteristics. Therefore, the efficiency of current Chinese medical text classification methods needs to be improved.

Dogra et al. [33] summarized some machine learning and deep learning algorithms used in the text classification field. Minarro Gimenez et al. [34] applied neural language models to the PubMed corpus, using skip-grams to represent the word representations in PubMed articles. TH et al. [35] used skip-gram and continuous bag of words (CBOW) to test on 1.25 million PubMed articles, verifying the performance of skip-gram and CBOW on large-scale corpora.

Table 1. Comparison of major works in the field of text classification.

Title	Contribution	Limitation
A Bayesian Classification Approach Using Class-Specific Features for Text Categorization [27]	Propose a Bayesian classification approach for automatic text categorization using class-specific features.	Only four classifications are supported, and the extended classification requires reconstruction of the model, the model has poor scalability.
Convolutional Neural Networks for Sentence Classification [28]	The structure of TextCNN is simple, the training speed is fast, and the classification effect is good, avoiding the process of feature selection in traditional models.	The convolution and pooling operations result in the loss of input information (e.g., lexical order, position), which makes it difficult for the model to learn negation, antonymy and other information in the text sequence.
Recurrent neural network for text classification with multi-task learning [29]	Propose the concept of shared layer, the model supports multi-task learning and pre-training.	Sentence-level corpus classification does not work well.
BertGCN: transductive text classification by combining GNN and BERT [45]	BertCGN combines large scale pre-training and transductive learning for text classification, which have a good performance on wide range of dataset.	Only using documents statistics to build graph structure, it might be sub-optimal.
ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations [46]	Propose a pre-training model for Chinese text, ZEN provides an alternative way of learning larger granular text.	The efficiency of training is low and it needs to be combined with other models for collaborative learning.

https://doi.org/10.1371/journal.pone.0282824.t001

However, the training speed of the CBOW model is slow. Inspired by the FastText [36] model, some scholars proposed some models such as CNN-LSTM [37] and DC-LSTM [38], and they used FastText to calculate word vectors and used LSTM model to solve the difficulty of information transmission caused by long sequence data. Edara et al. [31] analyzed moods of various cancer affected patients by collecting tweets from different online cancer-supported communities. They used several text mining and machine learning strategies to perform sentiment analysis on distributed LSTM framework and ran it in three different corpora, and the results showed that they were superior to some machine learning methods such as LDA and SVD in accuracy and efficiency. Srinivasu et al. [39] used recurrent neural networks such as LSTM and GRU to predict type 2 diabetes. At the same time, some scholars introduced the attention mechanism [40] into the application fields such as sentiment analysis and text classification, and the results indicate that attention mechanism is of great significance. Attention mechanism extracts meaningful words and sentences, then aggregates these representations to form sentence vectors [41], which is beneficial to the capture of key text features. Zhang et al. [42] proposed a label-based attention for hierarchical multilabel text classification neural network. The model extracted important information from labels at different levels. Lin et al. [43] proposed BertGCN, a model that combined large-scale pretraining and transductive learning for text classification. They used graph to model relationships between different samples from the whole corpus to utilize the similarity between labeled and unlabeled documents, and used GNNs to learn their relationships. However, BertGCN model only used document statistics to build the graph, and it is suboptimal compared with other models. Vulli et al. [44] introduced a novel method for the automated diagnosis and detection of metastases from whole slide images using the Fast AI framework and the 1-cycle policy. Table 1 show the contribution and limitation of some text classification methods.

Text classification algorithms based on deep learning mainly focus on the design of classifiers, such as local classifiers, global classifiers or their combination classifiers. Especially, in the field of Chinese medical text classification, there are two bottlenecks as follows:

1. Chinese medical text has a wide range of topics, widespread ambiguity, sparse features, and contain temporal features, which leads to low classification efficiency and a sharp drop in accuracy.

2. Chinese medical text contains a large number of medical nomenclatures with complex semantics, and hospital departments have different classification methods, so it is difficult to recommend appropriate departments for patients.

To overcome these limitations which in the current medical text classification approaches, this work mainly studies the information extraction and classification of Chinese medical text using dual-channel mechanism, so that our model takes the temporal and context factors into account, and effectively extracts text features from different granularities. The main contributions of our work are as follows:

1. We use pre-trained word vectors to encapsulate and represent Chinese medical text, so effectively compressing the dimensionality of feature vectors and alleviating data sparsity.
2. We propose a dual-channel Chinese medical text classification model with attention mechanism. The two channels are used to extract text information of different granularities, thereby extracting the temporal and contextual features of Chinese medical text.
3. The experimental results on CMDD and webMedQA datasets show that the proposed model has a great improvement in accuracy, recall, F1-score and balanced accuracy compared with state-of-the-art models.

The rest of this paper is organized as follows: the second section discusses the background of text classification and related research, the third section introduces our methodology and the framework of the proposed model, the fourth section analyzes the experimental results, and the fifth section summarizes our work and provides readers with hints on future directions to extend the proposed models.

Background and related work

Word embedding

The traditional vector space model (VSM) leads to a high dimension vector, and the features are independent of each other, which is inconsistent with the requirement that long text needs to be associated with the context. Bengio [47] proposed neural network language model (NNLM) and put forward the distributed representation of words in text for the first time. The distributed representation of words, that is, word vectors, can make the spatial distance between semantically similar words closer, thus containing more semantic information. The lower dimensions made it possible to train larger corpus and easier to solve the problem of dimensional disaster. Later, some tools such as Word2Vec [48], GloVe [49], FastText [36] for training word vectors appeared. This work uses the Word2Vec model to construct word vectors. Word2Vec includes two diametrically opposed models, skip-gram, and CBOW. The CBOW model trains the contextual word vector and outputs the word vector of the specific words. Usually, the department classification of a hospital is shown in Fig 1, and we utilize the CBOW method to encode the name of the department. The structure of the CBOW is shown in Fig 2.

GRU

The GRU [50] is a kind of Recurrent Neural Network (RNN), the GRU structure is a gated network structure. GRU solves the problem of gradient disappearance in long-term memory and backpropagation. Compared with LSTM, its structure is simpler, it only has two gates: update gate and reset gate. Therefore, less parameters need to be trained and calculated, and the training efficiency is higher. It is widely used in natural language processing tasks. The GRU structure is shown in Fig 3.

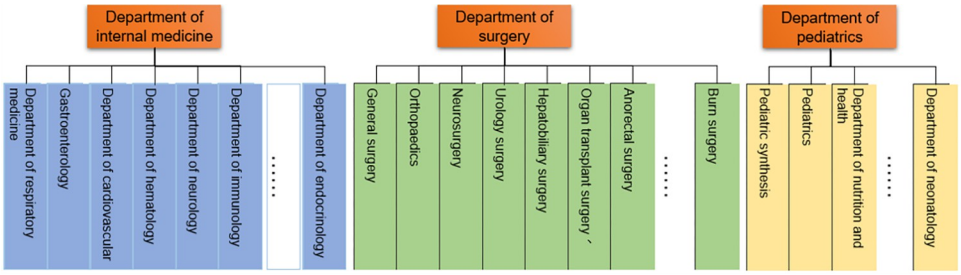

Fig 1. Hospital departments. Different hospitals have different department structures, Fig 1 shows some main departments in a hospital.

https://doi.org/10.1371/journal.pone.0282824.g001

The GRU model uses Eqs 1–4 to update the parameters.

$$Z_t = \sigma(w_z x_t + U_z h_{t-1}), \tag{1}$$

$$r_t = \sigma(w_t x_t + U_t h_{t-1}), \tag{2}$$

$$\tilde{h}_t = \tanh(w x_t + U(r_t \circ h_{t-1}), \tag{3}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t, \tag{4}$$

where z_t is the update gate of the GRU, r_t is the reset gate of GRU. The update gate controls the degree which the state information of the previous moment is transmitted to the current

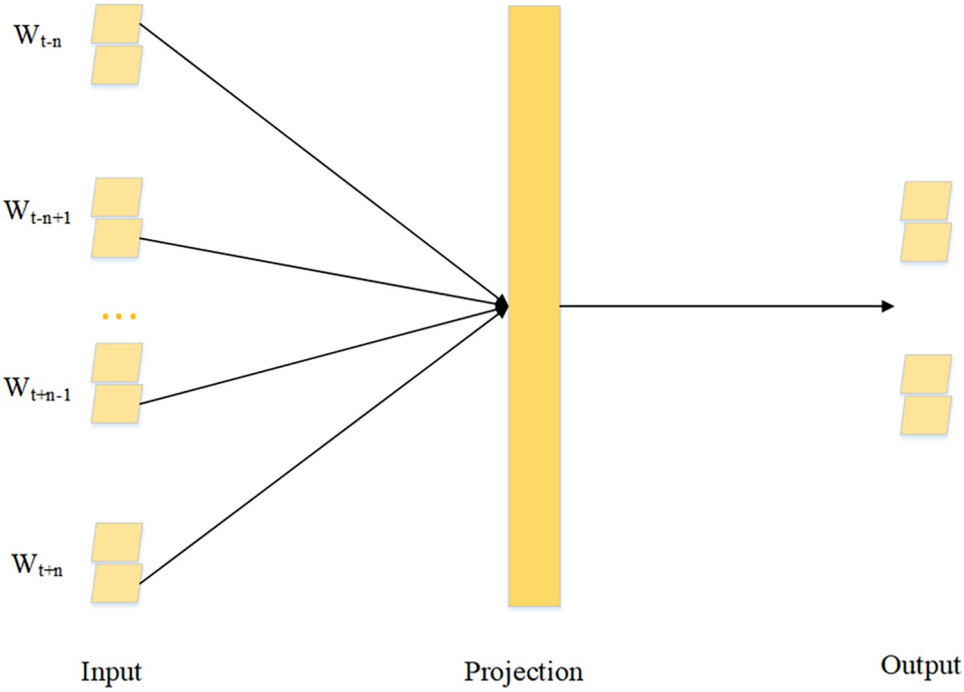

Fig 2. The structure of CBOW. CBOW is a neural network with a 3-layer structure, including input layer, hidden layer and output layer, and performs vector projection operations on the hidden layer. CBOW predicts the central word through the context.

https://doi.org/10.1371/journal.pone.0282824.g002

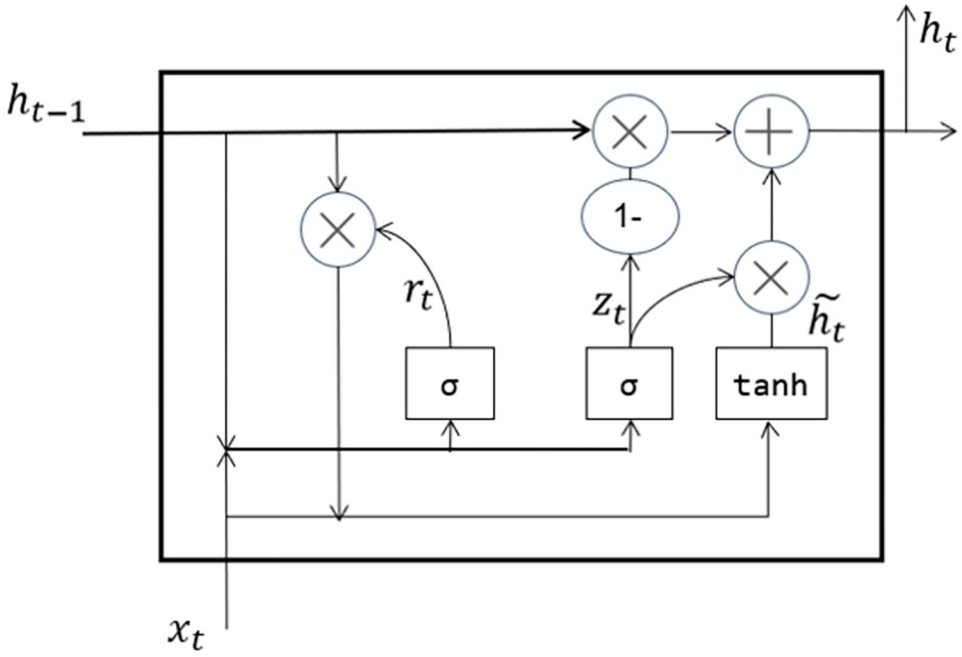

Fig 3. The structure of GRU. GRU only contains update gate and reset gate, and the structure is simpler. The update gate is used to control the degree to which the state information of the previous moment is brought into the current state. The larger the value of the update gate, the state information of the previous moment is brought into more. The reset gate controls how much information from the previous state is written into the current candidate set $\tilde{h}_t$.

https://doi.org/10.1371/journal.pone.0282824.g003

moment, and the reset gate controls the degree which the state information of the previous moment is forgotten. w_z, w_t and w are weights of GRU. GRU obtains two gated states through the hidden state information h_{t-1} at the previous moment and the node inputs x_t at the current moment. After obtaining the gated state, the reset gate obtains forgotten state, then fuses x_t at the current moment and activates it through the nonlinear function tanh, finally the update gate selects and memorizes the input of the current node.

The GRU is calculated from the front to the back according to the natural language order, only considering the impact of the above words in the text and the overall text semantics, ignoring the impact of the following words on the above and the overall semantics. Therefore, the bidirectional GRU (BiGRU) simultaneously propagates the calculation processing for the text input in positive order and reverse order, and integrates the corresponding position output, so that the text semantic information is more comprehensive and accurate.

Attention mechanism

The attention mechanism [51] was first proposed by the Google Mind team and applied to the field of image processing. It can be used by the multilayer structure of the network model for a single feature of an input sequence. However, in some scenarios, different models focus on the input sequence differently, that is, multiple features represent different aspects of the input sequence. The attention mechanism assigns different weight to these different representations. The more important the information, the higher the weight assigned, then greater influence on the classification. On the contrary, other information representations can be ignored to reduce the noise and redundancy of the input sequence to achieve better classification. The calculation of the attention mechanism can be divided into two steps: the first step is to calculate

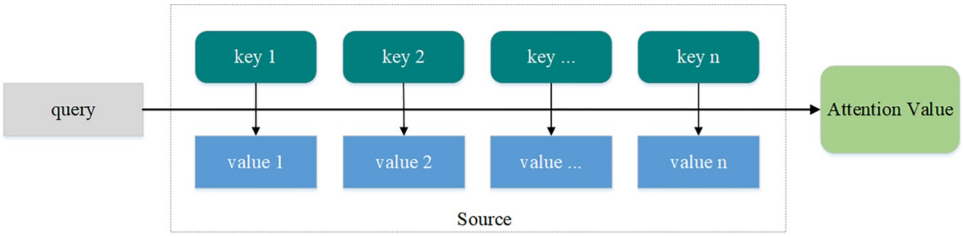

Fig 4. Attention mechanism. Attention is a set of attention distribution coefficients, and the attention value is obtained by using the attention function. The elements in Source are imagined as a series of <key, value> pairs. At this time, specify a query in the target, and calculate the similarity or correlation between query and each element to obtain the weight coefficient of each key corresponding to value, and then the value is weighted and summed to get the final attention value.

https://doi.org/10.1371/journal.pone.0282824.g004

the attention distribution in all input sequences, and the second step is to calculate the weight of the input information according to the attention distribution. The attention mechanism can essentially be described as a map of a query to a series of key-value pairs, as shown in Fig 4.

Methodology

The Dual-channel CNN-BiGRU-Attention (Dual-CGA) model proposed in this paper is a hierarchical hybrid text representation model. The overall structure is as shown in Fig 5. It is mainly composed of several parts: input embedding layer, convolutional and feedforward neural network channel layer, BiGRU and attention channel layer, and Softmax classification output layer. The model adopts a dual-channel mechanism. One channel uses convolution operations to obtain local information of text with different granularities, and the other

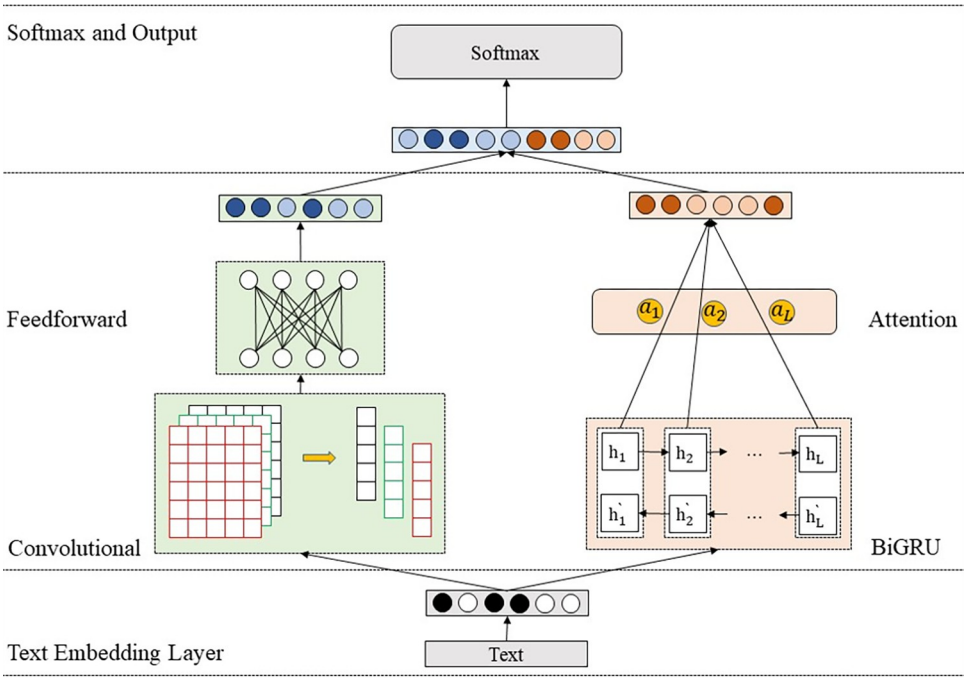

Fig 5. The framework of Dual-CGA. The model includes dual channels, four-layer structure. The dual channels are located in the convolutional and feedforward neural network channel layer. This layer is the core of the entire architecture, and integrated dual channels process text features of different granularities.

https://doi.org/10.1371/journal.pone.0282824.g005

channel uses BiGRU and attention mechanism to obtain global contextual semantic information. Finally, a full connection layer and Softmax classification are used to determine the classification of the outpatient department.

Medical text embedding layer

The medical text embedding layer aims to map the input text into a vector matrix composed of word vectors. The model uses the CBOW mode of Word2Vec to obtain the word vector as the representation element of the text vector. CBOW model predicts the probability of the target words through the window context and finally updates the parameter matrix, that is, the trained word vector representation. The word vector obtained by Word2Vec is generally from 100 to 300 dimensions, which is smaller than the one-hot coding dimension, so as to effectively avoid dimension disaster and data sparsity. Moreover, the word vector obtained by text training can better represent the semantic between words.

In this paper, word vectors are used in two ways: loading pre-trained word vectors and randomly initializing word vectors. The first step of both methods is to construct the vocabulary corresponding to the training set and then map the word sequence into a digital index sequence. By loading the pre-trained word vectors, we traverse and extract the word vector matrix corresponding to the vocabulary and generate word vector matrix by the text numerical index during the training stage. The strategy of randomly initializing word vectors is to convert text-digit sequences into randomly initialized word vectors trained with the model in the training stage, without loading pre-trained word vectors. The generation of the input text vector representation matrix W has the following steps:

Step 1: Segment each piece of data in the training set and construct a vocabulary list.

Step 2: For each piece of text in the training set and the test set, map the word segmentation result to the corresponding digital index of the vocabulary in which the index of words that have not appeared in the pre-trained vocabulary are filled with -1, and the word vector corresponding to the -1 index uses random initialization processing.

Step 3: Data length normalization operation, according to the data statistic result, the length of normalized value is set to 100, the excess part is truncated, the data with insufficient length is filled with 0, and the word vector corresponding to 0 index is initialized with all 0.

Step 4: Load the pre-trained word vectors, the index corresponding to the data is converted into word vector to obtain the word vector matrix W.

Through the above steps, the word vector matrix representation W is finally obtained, and W is used as the input of the model.

Convolutional and feedforward neural network channel layer

After the processing of the text embedding layer, the matrix representation of the word vector is obtained, but the word vector depends on the effect of word segmentation. Especially in Chinese medical text, there are many medical nomenclatures and terms. Poor word segmentation brings ambiguity and errors in text presentation into text encoding, resulting in greater errors in the next task. In addition, adjacent words in Chinese text are related, and the combination of words forms a structure containing more important contextual semantic information, such as subject-predicate, verb-object. The overall meaning of this structure can better represent the text. The Dual-CGA model uses convolution kernels to perform convolution operations on text according to different length of text, similar to n-grams, the model learns the semantic information contained in text of different granularities and covers multiple words of different length, thereby improves text semantics representational capabilities and performance on downstream text classification tasks.

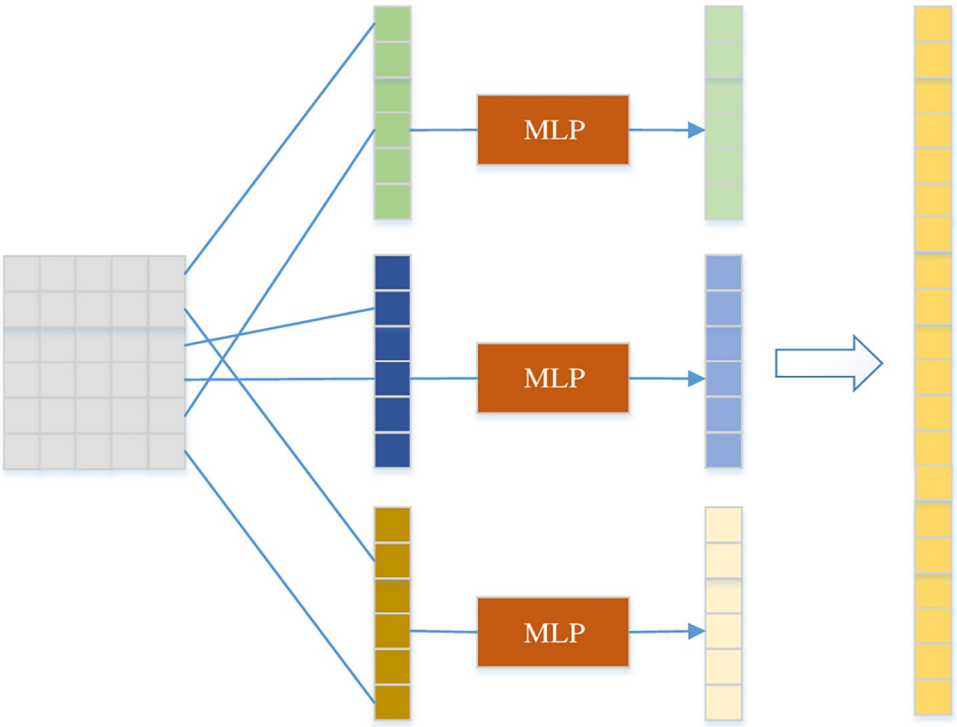

Fig 6. The structure of convolution and feedforward neural network channel layer. This layer firstly computes feature vectors by a convolutional structure consisting of three filters of different sizes, and then global information is obtained by MLP.

https://doi.org/10.1371/journal.pone.0282824.g006

The dimension of the word vector in this layer is 300 dimensions, and the convolutional neural network uses three filters, the sizes of which are 3*300, 4*300, and 5*300. Usually, each filter contains several feature maps. The filter only outputs a feature map vector, and then each feature map passes through the MLP layer. The MLP layer is equivalent to a matrix mapping with a global receptive field. It assigns weights to the results of convolution operations with different granularity and dynamically adjusts the parameters of the mapping matrix with network training, finally achieving the purpose of allocating high weight to important granularity information. The channel layer structure is as shown in Fig 6. This residual connection and hierarchical normalization operation can well prevent overfitting and reduce the occurrence of gradient disappearance during backpropagation, thus making Dual-CGA more robust.

The calculation process of this layer is as shown in Eqs 5–9,

$$c_3, c_4, c_5 = Conv(x), \tag{5}$$

$$c_3^{'} = f(w_3 \cdot c_3 + b_3), \tag{6}$$

$$c_4^{'} = f(w_4 \cdot c_4 + b_4), \tag{7}$$

$$c_5^{'} = f(w_5 \cdot c_5 + b_5), \tag{8}$$

$$c = [c_3^{'}, c_4^{'}, c_5^{'}]. \tag{9}$$

Among them, Eq 5 is convolution operation, which generates three feature vectors: c_3, c_4, c_5. Then input the feature vectors into the MLP network respectively and obtain the hidden layer output $c_3^{'}$, $c_4^{'}$, $c_5^{'}$, where w_3, w_4, w_5 and b_3, b_4, b_5 are the parameter matrix of the MLP network layer and bias vector. c is the concatenated long vector, that is, the text feature representation of the convolutional and feedforward neural network channel layer.

BiGRU and attention channel layer

In order to obtain more comprehensive text information, the model uses the BiGRU network to extract text information. The BiGRU network can extract the sequence features of the text sequence in the forward and reverse directions respectively. At the same time, the influence and relationship of context are considered; then, the text feature representation vector of the channel is obtained. The structure of BiGRU and attention channel layer is as shown in Fig 7.

$h^l_t = LSTM^l(x_t, h_{t-1})$ is the feature output of the forward hidden layer and $h^r_t = LSTM^r(x_t, h_{t-1})$ is the output of the reverse hidden layer. The hidden layer features in both directions are spliced together, that is, $h_t = [h^l_t, h^r_t]$, to get the output of this layer.

For the RNNs, the final output of the forward and reverse GRU network is spliced together and directly used as the text representation as the input of the text classifier. This method is relatively simple. Although the LSTM and GRU networks can process longer sequences, they can also learn longer text contextual features, there is still information attenuation. For classification tasks, this feature attenuation is fatal, resulting in low classification effects. In order to address this problem, the literature [34] proposed another way: splicing the output of the hidden layer at each moment together and then using the maximum pooling to obtain the most important feature representation as the input of the text classifier, which highlighted the most important features in the text. This method highlighted the most important features in the text, but ignored other secondary important features. In order to take into account the

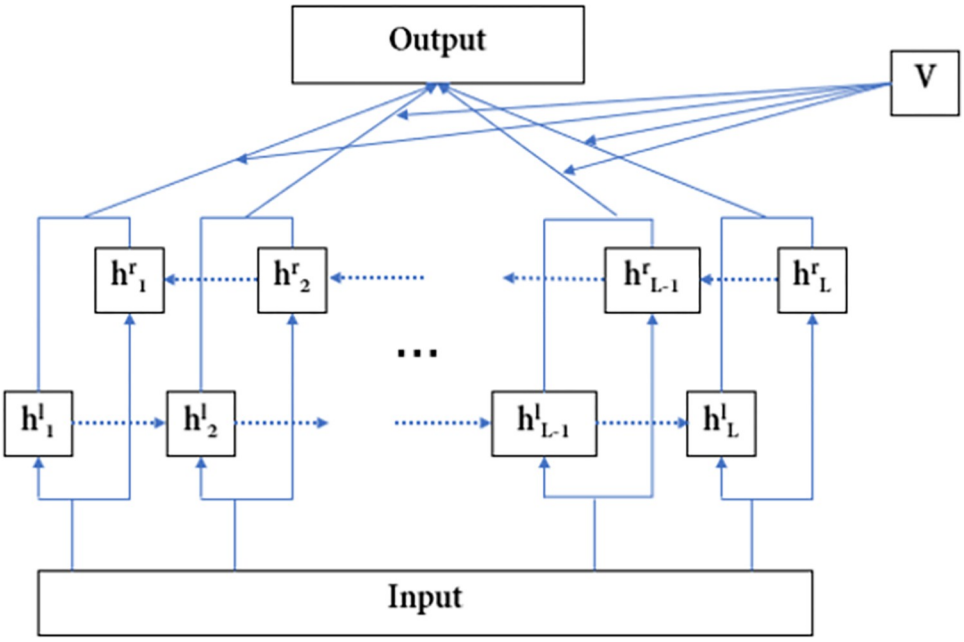

Fig 7. The structure of BiGRU and attention layer. This layer is composed of BiGRU and attention mechanism, which is used to fully obtain the long text context, and evaluate the importance of BiGRU output at each moment through the attention mechanism.

https://doi.org/10.1371/journal.pone.0282824.g007

characteristics of these different importance levels, our model introduces the attention mechanism to measure the importance of the output of GRU network at each time, so as to calculate the weight of the output at each time. The calculation method is as shown in following equations:

$$\mathrm{s}(x_i, \mathrm{q}) = x_i^T q$$

$$\mathrm{a}_i = softmax(s(x_i, q)) \tag{10}$$

$$V_a = \mathrm{attention}(\mathrm{V}) = \sum_{i=1}^{N} a_i V_i,$$

$$e_t = u_a^T \tanh(V_a \cdot h_t + b_a), \tag{11}$$

$$a_t = \frac{\exp(e_t)}{\sum_{i=1}^{L} \exp(e_i)}. \tag{12}$$

Eq 10 represents the computing method of the attention mechanism, Eq 11, e_t stands for h_t corresponding hidden layer output, V_a is the attention weight matrix, u_a^T is the random initialization vector, b_a is the bias. a_t in Eq 12 represents the normalized output of each hidden layer e_t, that is, the weight of the state output h_t at time t. After calculating the weight, the final representation can be obtained, as shown in Eq 13,

$$v_t = \sum_{t=1}^{L} a_t \cdot h_t, \tag{13}$$

The text feature representation of the final Bi-GRU and attention channel layer is recorded as v_t.

Softmax and classification output layer

The text feature representationr isis fused and spliced into a long vector to obtain the final text feature representation, denoted as o, as shown in the following equation:

$$o = c \oplus v, \tag{14}$$

The final text feature representation is input into the Softmax classifier, and gethe results is the probability distribution of the categories. The classification with the largest probability is selected as the final classification result.

Experiments

In order to verify the reliability of Dual-CGA model, the complete contrast experiments are designed. The two datasets of the Chinese medical dialogue data (CMDD) and webMedQA are used to compare the Dual-CGA model with other baseline models.

Datasets and hyperparameters

We conduct experiments on two Chinese medical public datasets of different scales, among which the larger dataset is CMDD, which is a public medical question answering dataset. The dataset contains 6 folders corresponding to 750,000 question-answer pairs related to 6

ID	37408749
Label	1
Category	内科
Translation	Internal Medicine
Question	早晨起来胃和肚子一起疼,拉稀,吃的早饭全早晨起来胃和肚子一起疼,拉稀,吃的早饭全吐了,吃了点氟哌酸肚子不疼了但是胃还一阵一阵的疼。
Translation	In the morning stomach and stomach ache together, diarrhea, eat breakfast all morning stomach and stomach ache together, diarrhea, eat breakfast all vomited, eat a little norfloxacin stomach pain but the stomach is still a burst of pain.
Answer	就你的描述考虑有胃肠炎的可能建议你服用胃必妥治疗。指导意见：建议你注重节制饮食，饮食上宜吃些清素、软烂、温热的食物，避免过早地进食肥肉、油炸食品、生冷坚硬的食品以及多纤维食物，如芹菜、黄豆芽、韭菜、蒜苔等。
Translation	Considering the possibility of gastroenteritis as described, you are advised to take metropin. Guidance: It is recommended that you pay attention to control diet, diet should eat some clear vegetarian, soft rotten, warm food, avoid eating fat meat, fried food, cold and hard food and fiber food, such as celery, bean sprouts, leek, garlic moss, etc.

Fig 8. A record sample of webMedQA. A record consists of 5 fields: ID, label, category, question and answer.

https://doi.org/10.1371/journal.pone.0282824.g008

different outpatient departments. Specifically, the number of detailed records corresponding to each department is as follows: andriatria (94,596 records), IM (220,606 records), OAGD (183,751 records), Oncology (75,553 records), Pediatric (101,602 records) and Surgical (115,991 records). Another dataset is the webMedQA, which is a question and answer dataset about Chinese medical questions collected from online healthcare websites. It contains more than 50,610 questions in 23 different categories,Each question has 1 positive and 4 negative answers. A sample is shown in Fig 8.

After data processing, The statistical results of the available data are described in Table 2.

As for how to define the confusion matrix under multi-classification, we take the CMDD dataset as an example, and define the labels corresponding to the 6 categories of the dataset as 0–5 in turn, and the TP, TN, FP and FN with category 0 are calculated as follows:

TP: the actual value is 0, and the predicted value is also 0;

FP: the predicted value is 0, the actual value is not 0;

FN: the actual value is 0, the predicted value is not 0;

TN: neither the actual value nor the predicted value is 0.

It is well known that the quality of hyperparameters will directly affect the training effect of the model, so it is important to choose a series of optimal hyperparameters. The settings of the hyperparameters are shown in Table 3.

This experimental environment is based on 64 bit Anaconda as the modeling platform, and the operating system is Ubuntu 20. The experimental environment is shown in Table 4.

Baselines and evaluation

This paper conducts experiments on two datasets with 1 experimental group and 8 control groups. The experimental group was designed based on the Dual-CGA model, while the control groups used 3 basic text classification models, Naïve Bayes [27], SVM [26] and CNN [28],

Table 2. Statistics of datasets.

Dataset	Categories	Records	Average length of records
CMDD	6	752,236	389.43
webMedQA	23	50,610	86.68

https://doi.org/10.1371/journal.pone.0282824.t002

Table 3. Hyperparameters.

Parameters	Values
Vector dimensions	300
Batch size	128
Kernel size	3/4/5
Learning rate	1e-3
Dropout	0.25
Optimizer	Adam

https://doi.org/10.1371/journal.pone.0282824.t003

as well as RCNN [30], BiGRU [50], BiLSTM, Transformer [45], ZEN [46] and InducT-GCN [52] models.

- Naïve Bayes: Naïve Bayes assumes the conditional independence of conditional probability distribution, that is, ignores the relationship between features, making each feature a separate hypothesis. When using Naïve Bayes for text classification, it is more effective in the case of less data, but it is sensitive to input data, especially Chinese medical text.
- SVM: a benchmarked supervised-learning approach that can handle nonlinear classification. The aggregation words proposed by VSM are embedded as features.
- CNN: a classic variant of convertional deep neural networks that implements convolutional filters with learned weights and bias.
- BiGRU: BiGRU can calculate the input sequence from front to back, which extract the features more accurately and grasp the context relationship of medical text.
- BiLSTM: a bidirectional LSTM developed by [53] with a concatenated layer of one forward LSTM and one backward LSTM and a hidden state.
- Transformer: Transformer can train all words in parallel at the same time, which greatly speeds up the computational efficiency. Moreover, Transformer adds location embedding to help the model understand the order of language. The structure of Transformer includes self-attention and full connection layer.
- ZEN: The model [46] is a BERT-based Chinese text encoder enhanced by N-gram representations that take into account different combinations of characters during training.
- InducT-GCN: The model [52] is a inductive graph-based text classfication framework without using extra resources.

Table 4. Experimental environment.

Hardware and software	Values
CPU	Intel i9 10900K
RAM	128G
GPU	GeForce RTX 3090
Display RAM	24G
OS	Ubuntu 20
Development language	Python 3.9
Python distribution platform	Anaconda 64-Bit
ML Platform	TensorFlow 2.2
Chinese word segmentation tool	Jieba 0.42.1
Scientific computing package	NumPy 1.19.5

https://doi.org/10.1371/journal.pone.0282824.t004

Table 5. Evaluation indicators.

Name	Description
Accuracy	$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
Recall (R)	$R = \frac{TP}{TP+FN}$
Precision (P)	$P = \frac{TP}{TP+FP}$
F1-score	$F1 - score = \frac{2*P*R}{P+R}$
Balanced Accuracy (BA)	$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$

https://doi.org/10.1371/journal.pone.0282824.t005

We use some popular evaluation indicators in machine learning to evaluate the performance of the models, including accuracy, recall, F1-score and balanced accuracy (BA), as shown in Table 5.

Accuracy represents the overall prediction ability of the deep learning model. True positive (TP) and true negative (TN) measure the ability of classifier models to predict the correct department to recommend for the patients. False positive (FP) and false negative (FN) identify the number of incorrectly predictions generated by the models. Recall measures sensitivity of the classifier models. F1-score keeps the balance between precision and recall. It's often used when class distribution is uneven, but it can also be defined as a statistical measure of the accuracy of an individual test. Balanced accuracy is useful for multiclass classification, BA is the average of recall obtained in each class.

Results and analysis

In order to avoid model overfitting as much as possible, the optimization method we choose is the dropout method. The mechanism of this method is to delete some nodes during propagation, so that we can reduce the network complexity, and the established neural network model will not become overfitting. The experimental results are shown in Tables 6 and 7.

It can be seen from Tables 6 and 7 that the Dual-CGA model proposed in this paper is superior to other models such as CNN, RCNN, BiLSTM, BiGRU, Transformer and InducT-GCN in terms of accuracy, recall, F1-score on CMDD dataset but slightly lower than ZEN model in terms of F1-score. However Dual-CGA shows an advantage on the webMedQA dataset, which is much better than other comparison models in terms of accuracy, recall and F1-score. Figs 9 and 10 describe the comparison between the Dual-CGA model and other baseline models. It is observed that the performance of our model is better than other models.

Table 6. Experimental results on CMDD.

Models	Accuracy (%)	Recall (%)	F1-score (%)	BA (%)
NB	82.78	82	82	83.92
SVM	81.63	79	81	82.96
CNN	87.33	86.33	86.76	87.96
RCNN	88.46	87.9	88.02	88.98
BiLSTM	88.49	87.92	87.99	89.02
BiGRU	88.85	88.17	88.38	89.35
Transfomer	88.15	87.94	87.66	88.72
ZEN	90.65	90.31	90.68	91.06
InducT-GCN	82.03	82.04	82.03	83.4
Dual-CGA	**90.8**	**90.53**	**90.51**	**91.16**

https://doi.org/10.1371/journal.pone.0282824.t006

Table 7. Experimental results on webMedQA.

Models	Accuracy (%)	Recall (%)	F1-score (%)
NB	67.32	37	37
SVM	64.2	36	37
CNN	73.13	33.95	34.81
RCNN	68.25	30.93	32.74
BiLSTM	69.91	34.6	33.5
BiGRU	71.39	36.72	36.19
Transfomer	71.69	37.3	39.2
ZEN	23.45	16.66	6.33
InducT-GCN	61.72	61.72	61.72
Dual-CGA	**75.11**	**42.2**	**45.06**

https://doi.org/10.1371/journal.pone.0282824.t007

Compare Figs 9 and 10, it can be seen that the results of the Dual-CGA model are quite different. In fact, not only the Dual-CGA, but also other models, the main reason lies in the dataset itself. For example, the text average length of webMedQA is only 86.68 characters, and its average length is only one-fourth of the CMDD dataset, resulting in the number of extracted features is small. In addition, the webMedQA dataset has only more than 40000 records, but the amount of its catalogs is 1.83 times great than that of CMDD dataset, resulting worse experimental results on webMedQA. The InducT-GCN model is a corpus-level text classification model, which has certain advantages in short text classification. This also points out the direction for our future research, that is, to improve classification effect of short text corpus. Moreever, Dual-CGA has better performance than baseline models, among them, it is the model with the shortest traning time.

Ablation studies

We conduction ablation studies on CMDD dataset, we successively adjusted three different learning rates, then removed the attention mechanism, and adjusted the kernel size. In these different hyperparameter states, there is a significant impact on the execution results of the model, as shown in Table 8.

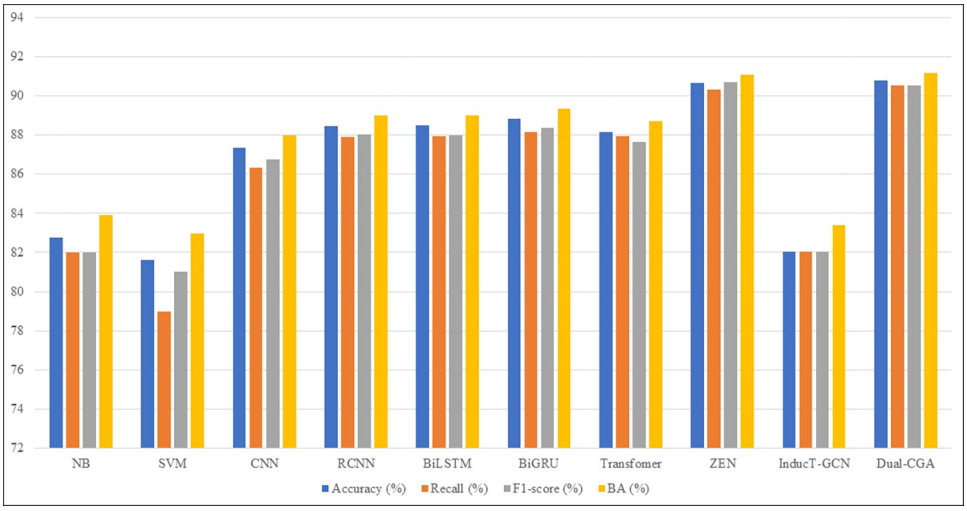

Fig 9. Experimental results on CMDD dataset. The experimental results in the CMDD dataset show that Dual-CGA is almost superior to the baseline models in the four indicators of accuracy, recall, F1-score and BA.

https://doi.org/10.1371/journal.pone.0282824.g009

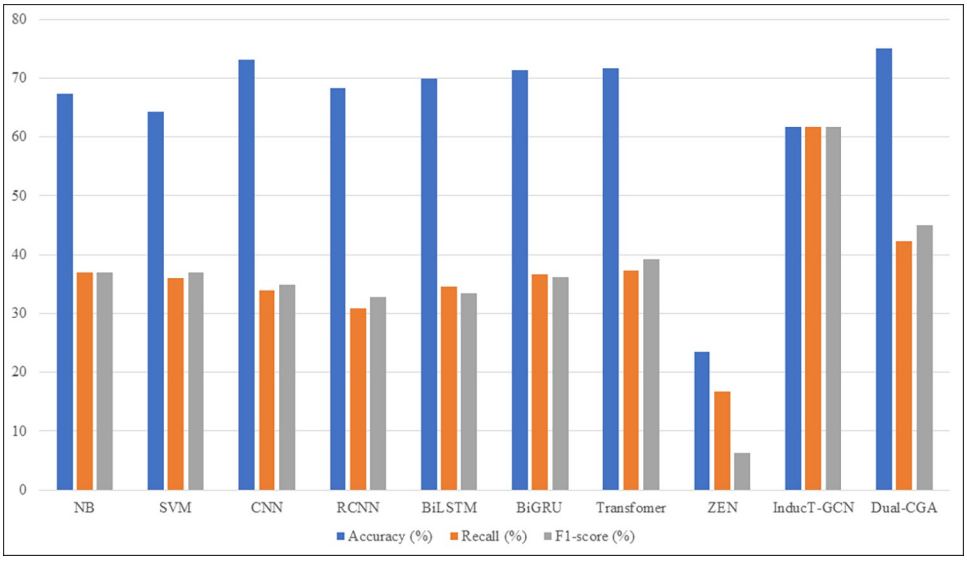

Fig 10. Experimental results on webMedQA dataset. The experimental results on the webMedQA dataset show that Dual-CGA also has certain advantages in indicators such as accuracy, recall, and F1-score, but it is inferior to inductT-GCN in terms of recall and F1-score, because the average text length of the dataset is shorter, the ability of BiGRU to extract short text features is not as good as that of the pre-trained structure model.

https://doi.org/10.1371/journal.pone.0282824.g010

Fig 11 shows that the performance of the model's accuracy and loss values during the training phase is basically consistent with the performance during the test phase, and the model does not have the problem of overfitting.

We adjust different learning rates and compare the performance of the loss value under the four different values of the learning rate of 1e-2, 1e-3, 1e-4 and 1e-5. Fig 12 clearly shows that when the learning rate is 1e-3, the loss value is lowest.

Conclusions

This paper solves a problem of accurately recommending medical departments for patients based on their symptom descriptions. We focus on the structural features of Chinese medical text and propose a Chinese medical text classification model based on multi-channel and attention mechanism, namely Dual-CGA. The model adopts a dual-channel method including convolutional neural network and BiGRU, which is used to identify different granularities and medical nomenclature of Chinese medical care. At the same time, an attention mechanism is introduced to increase the weight of important word vectors, and have a good classification effect. A large number of experimental results show that our model has good performance in text classification, the accuracy, recall and F1-score are respectively 10.65%, 8.94% and 11.62%

Table 8. Ablation tests on different hyperparameters.

Hyperparameters	Accuracy	Recall	F1-score
Without attention	0.8985	0.8914	0.8931
Kernel size (2/3/4)	0.9003	0.8904	0.8937
lr = 1e-4	0.8950	0.8868	0.8891
lr = 1e-2	0.8756	0.8651	0.8693
lr = 1e-5	0.8585	0.8466	0.8495

https://doi.org/10.1371/journal.pone.0282824.t008

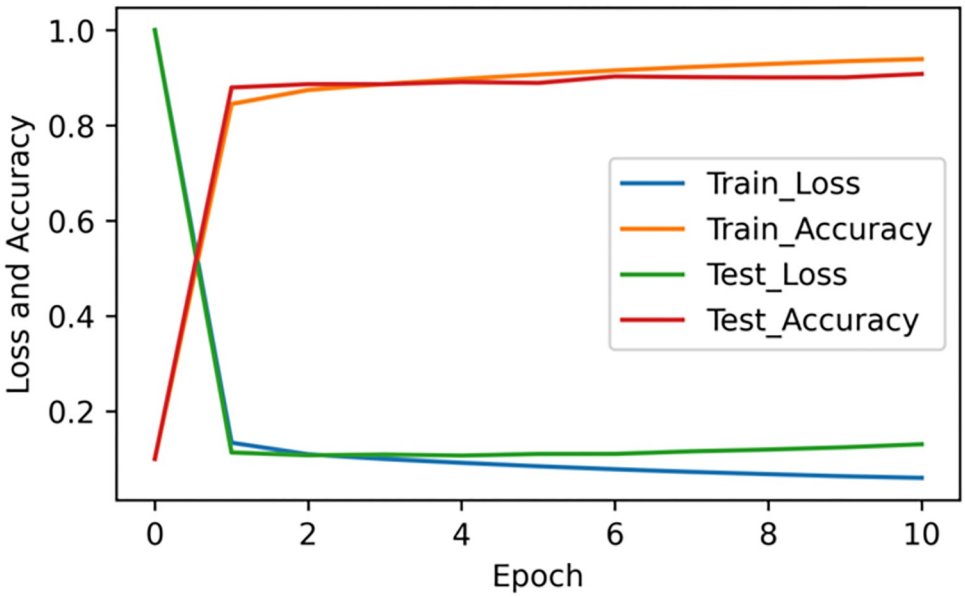

Fig 11. Loss associated with batches processed before fine-tuning the model. The performance of the Dual-CGA model in the training phase is consistent with that in the test phase.

https://doi.org/10.1371/journal.pone.0282824.g011

higher than the baseline model on average. This work can recommend appropriate departments for patients, help hospitals effectively triage patients and improve patient experience.

In the future, we will further mine the characteristics of each outpatient department, because there are still many hidden patterns and rules to be discovered to achieve better classification results. Meanwhile, we also try to address performance issues with short text and apply Dual-CGA to other fields, such as sentiment analysis and disease prediction.

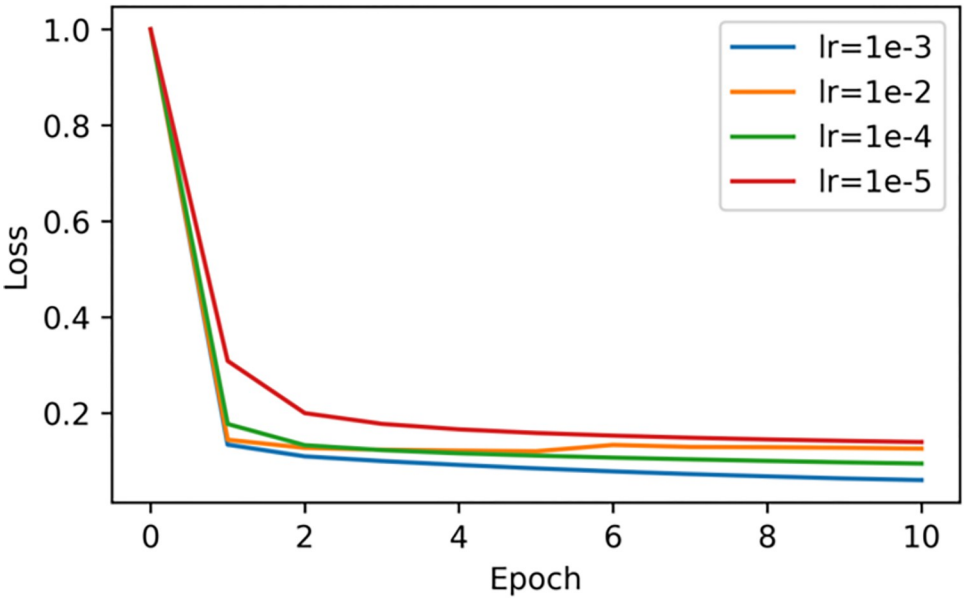

Fig 12. The loss value with different learning rate. Fig 12 shows that when the learning rate is 1e-3, the loss value is the smallest.

https://doi.org/10.1371/journal.pone.0282824.g012

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their valuable comments.

Author Contributions

Conceptualization: Shengbin Liang.

Data curation: Jiangyong Jin.

Formal analysis: Na Li.

Methodology: Yuying Zhang, Shengbin Liang.

Software: Xiaoli Li, Yuying Zhang, Jiangyong Jin, Fuqi Sun.

Validation: Xiaoli Li, Fuqi Sun.

Visualization: Xiaoli Li, Yuying Zhang.

Writing – review & editing: Shengbin Liang.

References

1. El-Sappagh S., Ali F., Abuhmed T., Singh J., Alonso J. M. Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing. vol. 512; no. 2022, pp. 203–224, 2022.
2. Ali F., El-Sappagh S., Islam S.M. R., Kwak D., Ali A., Imran M., et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion. vol. 63; no. 2020, pp. 208–222, 2020.
3. Ali F., El-Sappagh S., Islam S.M. R., Ali A., Attique M., Imran M., et al. An intelligent healthcare monitoring framework using wearable sensors and social networking data. Future Generation Computer Systems. vol. 114; no. 2021, pp. 23–43, 2020.
4. Alfian G., Syafrudin M., Ijaz M. F., Syaekhoni M. A., Fitriyani N. L., Rhee J. A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing. Sensors. vol. 18; no. 7, pp. 2183, 2018. https://doi.org/10.3390/s18072183 PMID: 29986473
5. Srinivasu P. N., SivaSai J. G., Ijaz M. F., Bhoi A. K., Kim W., Kang J. J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors. vol. 21; no. 8, pp. 2852, 2021. https://doi.org/10.3390/s21082852 PMID: 33919583
6. Wang Y., Wang L., Rastegar-Mojarad M., Moon S., Shen F., Afzal N., et al. Clinical information extraction applications: a literature review. Journal of Biomedical Informatics. vol. 77; pp. 34–49, 2018. https://doi.org/10.1016/j.jbi.2017.11.011 PMID: 29162496
7. Mykowiecka A., Marciniak M., Kupść A. Rule-based information extraction from patients' clinical data. Journal of Biomedical Informatics. vol. 42; no. 5, pp. 923–936, 2009. https://doi.org/10.1016/j.jbi.2009.07.007 PMID: 19646551
8. Kluegl P., Toepfer M., Beck P.-D., Fette G., Puppe F. UIMA Ruta: rapid development of rule-based information extraction applications. Natural Language Engineering. vol. 22; no. 1, pp. 1–40, 2016.
9. Zhu H., Ni Y., Cai P., Qiu Z., Cao F. Automatic extracting of patient-related attributes: disease, age, gender and race. Studies in Health Technology and Informatics. vol. 180; no. 1, pp. 589–593, 2012. PMID: 22874259
10. Shen F., Wang L., Liu H. Phenotypic analysis of clinical narratives using human phenotype ontology. Studies in Health Technology and Informatics. vol. 245; pp. 581–585, 2017. PMID: 29295162
11. Franois S., Erik S., Nicolas M., Hassina L. Gabriel Non-redundant association rules between diseases and medications: an automated method for knowledge base construction. Bmc Medical Informatics and Decision Making. vol. 15; no. 1, p. 7, 2015.
12. Kuo T. T., Rao P., Maehara C., Doan S., Chaparro J. D., Day M. E., et al. Ensembles of NLP tools for data element extraction from clinical notes. Amia Annu Symp Proc Hsu. pp. 1880–1889, 2016. PMID: 28269947
13. Savova G. K., Masanz J. J., Ogren P. V., Zheng J., Sohn S., Kipper-Schuler K. C., et al. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and

applications. Journal of the American Medical Informatics Association. vol. 17; no. 5, pp. 507–513, 2010. https://doi.org/10.1136/jamia.2009.001560 PMID: 20819853

14. Wu Y., Xu J., Jiang M., Zhang Y. A study of neural word embeddings for named entity recognition in clinical text. AMIA Annual Symposium proceedings/AMIA Symposium. vol. 1326, 2015.
15. Li R., Zhao H., Lin Y., Maxwell A., Zhang C. Multi-label classification for intelligent health risk prediction. Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine. pp. 986–993, Shenzhen, China, December 2017.
16. Li C., Konomis D., Neubig G., Xie P., Cheng C., Xing E. Convolutional Neural Networks for Medical Diagnosis from Admission Notes. 2017. https://arxiv.org/abs/1712.02768.
17. Kuo M.-H., Gooch P., St-Maurice J. A proof of concept for assessing emergency room use with primary care data and natural language processing. Methods of Information in Medicine. vol. 52; no. 01, pp. 33–42, 2013. https://doi.org/10.3414/ME12-01-0012 PMID: 23223678
18. Hsu W., Han S. X., Arnold C. W., Bui A. A., Enzmann D. R. A data-driven approach for quality assessment of radiologic interpretations. Journal of the American Medical Informatics Association. vol. 23; no. e1, pp. e152–e156, 2015. https://doi.org/10.1093/jamia/ocv161 PMID: 26606938
19. Yuan L. Recurrent neural networks for classifying relations in clinical notes. Journal of Biomedical Informatics. vol. 72; pp. 85–95, 2017. https://doi.org/10.1016/j.jbi.2017.07.006 PMID: 28694119
20. Li D., Liu S., Rastegar-Mojarad M., Wang Y. A Topic-Modeling Based Framework for Drug-Drug Interaction Classification from Biomedical Text. Amia Annu Symp Proc. vol. 2017; pp. 789–798, 2016. PMID: 28269875
21. Chen J., Druhl E., Polepalli Ramesh B., Houston T. K., Brandt C. A., Zulman D. M., et al. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. Journal of Medical Internet Research. vol. 20; no. 1, p. e26, 2018. https://doi.org/10.2196/jmir.8669 PMID: 29358159
22. Névéol A. and Zweigenbaum P. Making sense of big textual data for health care: findings from the section on clinical natural language processing. Yearbook of Medical Informatics. vol. 26; no. 01, pp. 228–234, 2017. https://doi.org/10.15265/IY-2017-027 PMID: 29063569
23. Chen J., Jagannatha A. N., Fodeh S. J., Yu H. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: adapted distant supervision approach. JMIR Medical Informatics. vol. 5; no. 4, p. e42, 2017. https://doi.org/10.2196/medinform.8531 PMID: 29089288
24. Wang Y., Ravikumar K. E., Rastegar-Mojarad M., Liu H. Leveraging word embeddings and medical entity extraction for biomedical dataset retrieval using unstructured texts. Database, vol. 2017, 2017. https://doi.org/10.1093/database/bax091 PMID: 31725862
25. Henriksson A., Kvist M., Dalianis H., Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. Journal of Biomedical Informatics. vol. 57; pp. 333–349, 2015. https://doi.org/10.1016/j.jbi.2015.08.013 PMID: 26291578
26. Yan J., Li J., Gao X. Chinese text location under complex background using Gabor filter and SVM. Neurocomputing. vol. 74; no. 17, pp. 2998–3008, 2011.
27. Tang B., He H., Baggenstoss P. M., Kay S. A bayesian classification approach using class-specific features for text categorization. IEEE Transactions on Knowledge and Data Engineering. vol. 28; no. 6, pp. 1602–1606, 2016.
28. Kim Y. Convolutional Neural Networks for Sentence Classification. 2014, https://arxiv.org/abs/1408.5882.
29. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning. 2016. https://arxiv.org/abs/1605.05101.
30. Lai S., Xu L., Liu K., Zhao J., Recurrent convolutional neural networks for text classification. Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence, pp. 2267–2273, Austin, TX, USA, 2015.
31. Edara D. C., Vanukuri L. P., Sistla V., Kolli V. K. K. Sentiment analysis and text categorization of cancer medical records with LSTM. Journal of Ambient Intelligence and Humanized Computing, pp. 1–17, 2019.
32. Tang X., Chen Y., Dai Y., Xu J., Peng D. A multi-scale convolutional attention based GRU network for text classification. Chinese Automation Congress, (CAC), pp. 3009–3013, 2019.
33. Dogra V., Verma S., Kavita P. Chatterjee, Shafi J., Choi J., et al. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. Computational Intelligence and Neuroscience, vol. 2022; no. 12, p. 1883698, 2022. https://doi.org/10.1155/2022/1883698 PMID: 35720939

34. Minarro-Giménez J. A., Marín-Alonso M. O., Samwald M. Exploring the application of deep learning techniques on medical text corpora. Studies in Health Technology and Informatics. vol. 205; pp. 584–588, 2014. PMID: 25160253

35. Muneeb T. H., Sahu S., Ashish A. Evaluating distributed word representations for capturing semantics of biomedical concepts. Proceedings of BioNLP. vol. 15; pp. 158–163, 2015.

36. Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification. 2017, https://arxiv.org/abs/1607.01759.

37. Zeghdaoui M. W., Omar B., Bentayeb F., Joly F. Medical-based text classification using FastText features and CNN-LSTM model. Proceedings of the International Conference on Database and Expert Systems Applications. vol. 12923; pp. 155–167, Virtual Event, Springer, Cham, Switzerland, September 2021.

38. Liang S., Chen X., Ma J., Du W., Ma H. An improved double channel long short-term memory model for medical text classification. Journal of Healthcare Engineering. vol. 2021; pp.6664893, 2021. https://doi.org/10.1155/2021/6664893 PMID: 33688423

39. Srinivasu P. N., Shafi J., Krishna T. B., Sujatha C. N., Praveen S. P., Ijaz M. F. Using Recurrent Neural Networks for Predicting Type-2 Diabetes from Genomic and Tabular Data. Diagnostics. vol. 12; no. 12, pp. 3067, 2022.

40. Li W., Qi F., Tang M., Yu Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. Neurocomputing. vol. 387; pp. 63–77, 2020.

41. Ren J., Wu W., Liu G., Chen Z., Wang R. Bidirectional gated temporal convolution with attention for text classification. Neurocomputing. vol. 455; no. 1, pp. 265–273, 2021.

42. Zhang X., Xu J., Soh C., Chen L., LA-HCN: label-based attention for hierarchical multi-label text classification neural network. Expert Systems with Applications. vol. 187; pp. 115922, 2022.

43. Lin Y., Meng Y., Sun X., Han Q., Kuang K., Li J., et al. BertGCN: transductive text classification by combining GNN and BERT. 2021. https://doi.org/10.48550/arXiv.2105.05727

44. Vulli A., Srinivasu P. N., Sashank M. S. K., Shafi J., Choi J., Ijaz M. F. Fine-Tuned DenseNet-169 for Breast Cancer Metastasis Prediction Using FastAI and 1-Cycle Policy. Sensors. vol. 22; no. 8, pp. 2988, 2022. https://doi.org/10.3390/s22082988 PMID: 35458972

45. Shaheen Z., Wohlgenannt G., Filtz E. Large scale legal text classification using transformer models. 2020. https://doi.org/10.48550/arXiv.2010.12871

46. Diao S., Bai J., Song Y., Zhang T., Wang Y. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. 2019. https://doi.org/10.48550/arXiv.1911.00720

47. Bengio Y. and Senecal J.-S., "Adaptive importance sampling to accelerate training of a neural probabilistic language model," IEEE Transactions on Neural Networks, vol. 19, no. 4, pp. 713–722, 2008. https://doi.org/10.1109/TNN.2007.912312 PMID: 18390314

48. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. 2013. https://doi.org/10.48550/arXiv.1301.3781

49. Pennington J., Socher R., Manning C. D. Glove: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543, (EMNLP), Doha, Qatar, October 2014.

50. Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. https://doi.org/10.48550/arXiv.1406.1078

51. Mnih V., Heess N., Graves A. Recurrent models of visual attention. Advances in Neural Information Processing Systems. vol. 27; pp. 2204–2212, 2014.

52. Wang Kunze, Soyeon Caren Han Josiah Poon. InducT-GCN: Inductive Graph Convolutional Networks for Text Classification. 2022. https://doi.org/10.48550/arXiv.2206.00265

53. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. vol. 18; no. 5–6, pp. 602–610, 2005. https://doi.org/10.1016/j.neunet.2005.06.042 PMID: 16112549