Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 6, 2024

ALBERT-Driven Ensemble Learning for Medical Text Classification

Yiru Cang[1], Wangying Yang[2], Dan Sun[3], Zhi Ye[4], Zitao Zheng[5]

[1]Northeastern University, Boston, USA

[2]University of Southern California, Los Angeles, USA

[3]Independent Researcher, Redwood City, USA

[4]Elevance Health, Indianapolis, USA

[5]Independent Researcher, New Jersey, USA

Correspondence should be addressed to Zitao Zheng; vzz29urz@gmail.com

Abstract: Health queries, as a specialized form of medical text, present unique challenges due to the presence of complex medical terminology, abbreviations, and linguistic features such as synonyms, antonyms, and polysemy. Traditional text classification methods often struggle with the intricacies of category labels, hierarchical relationships, and the scarcity of annotated data samples. This study presents an advanced medical text classification method utilizing the ALBERT pre-trained language model for health queries. We introduce the TLCM and TCLA models, which apply transfer learning and ensemble learning to enhance classification accuracy. By fine-tuning the ALBERT model and integrating CNN, Bi-LSTM, and Attention mechanisms, our models achieve approximately 91% in Precision, Recall, and Micro_F1, significantly improving upon traditional classification methods. This approach demonstrates the potential of pre-trained language models in medical text mining.

Keywords: Medical Text Classification, Pre-trained Language Models, Transfer Learning, Attention mechanisms

1. Introduction

Health queries [1], as a specific form of medical text, are characterized by complex medical jargon and a plethora of abbreviations, as well as phenomena common to natural language such as synonyms, antonyms, and polysemy. During text processing, challenges arise from the complexity of category labels and hierarchical relationships, the scarcity of effectively annotated data samples, and the high semantic similarity between different query texts. This makes the text classification task in this specific medical field more challenging.

This paper primarily explores the optimal learning model for small samples in a specific target domain, leveraging the general and powerful generalization capabilities of pre-trained language models. Based on the ALBERT [2] benchmark model framework, a medical text classification method based on transfer learning and ensemble learning is proposed, along with two variant models: the TLCM model and the TCLA model. The main work and innovations are as follows.

(1) The ALBERT pre-trained language model, which performs well in general domains, is introduced for dynamic word vector representation. The model fine-tuning technique is used to adjust the Embedding input layer structure, the multi-layer bidirectional Transformer structure, and the network structure of the downstream classification subtask. In this process, the Embedding input layer employs transfer learning methods to input health query descriptive texts at the character level for character vector representation.

(2) The original multi-layer bidirectional Transformer structure within the ALBERT model is transferred, and the trained output vectors are combined with multiple hybrid neural network modules such as CNN [3] structures, Bi-LSTM [4] structures, and Attention mechanisms for supervised ensemble training. The TLCM and TCLA model frameworks are proposed to further extract local information features and global structural information features from the text to construct classifiers.

(3) In the downstream task, a multi-label classification subtask is constructed, and two fully connected multi-layer perceptrons are designed to construct a text multi-label classifier. The context representation of the text is labeled using the cross-entropy [5] mechanism and sigmoid [6] activation function to achieve thematic classification of health query descriptive texts.

Experimental results show that in the multi-label classification task of health queries, the TLCM and TCLA models proposed in this paper have achieved around 91% in various evaluation metrics such as Precision, Recall, and Micro_F1, demonstrating good performance. They effectively address the shortcomings of traditional text classification algorithms in understanding the semantics of medical texts, having single category labels, and lower classification accuracy. Compared to the static word vector representation of traditional word2vec [7], the introduction of pre-trained language models significantly enhances algorithm performance and shows great promise in the field of medical text information mining.

2. Background

Medical texts, as a form of special and complex sequential information, present challenges due to their semantic sparsity and high dimensionality, as well as the multi-label classification requirements for identifying the thematic categories to which texts belong. This chapter proposes a text classification algorithm based on transfer learning and ensemble learning, which is tailored to address these challenges. The model's general framework and design philosophy are elaborated in detail, and based on the objectives of the multi-label classification task and the basic framework of the proposed model, the benchmark model is extended in two independent directions, leading to the proposal of two distinct variant models. The work presented in this paper is thoroughly introduced from the perspectives of mathematical theory, model framework, algorithmic process, and algorithmic description.

Let the entire dataset of health query texts be denoted as $w = \{(W_k, y_k')_{j=1}^{L}\}_{k=1}^{N} \in D$, $(W_k)_{k=1}^{N}$ represents the k-th health query description text in the sample, N represents the total number of texts, and $(y_k^f)_{j=1}^{L}$ represents the label categories corresponding to each health query description text, L represents the total number of labels, and each label $y_k^J \in \{0,1\}$ indicates whether the text belongs to the current label category, with "0" indicating "no" and "1" indicating "yes". This paper constructs a deep ensemble learning framework based on the ALBERT pre-trained language model to process medical texts. The specific research objective is to, given a health query text description $W = \{w_1, w_2, \cdots, w_n\}$ and a candidate label pool $Y = \{y_1, y_2, \cdots, y_L\}$,, search for and output the optimal label candidate answers, where n represents the length of the text. The mapping relationship between the health query description text and the labels is shown in Figure 1.

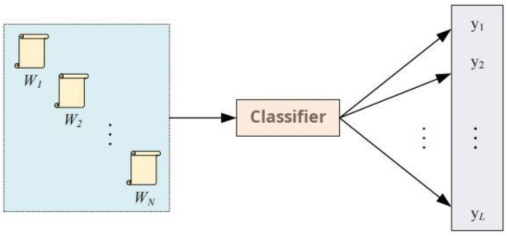

Figure 1. Text-Label Mapping Relationship

A health query description text is typically a sequence of arbitrary length, and predicting the multiple thematic categories involved in this health query text presents a challenge. The correct labels may not share lexical units directly with the text sequence; instead, they may only have some semantic association. Additionally, health query description texts may contain a large amount of irrelevant information, noisy data quality, and imbalanced label distribution, which poses significant challenges to the classification task. In this paper, we propose a transfer learning [8] and ensemble framework [9] for multi-label medical text classification. Based on the version of the ALBERT pre-trained language model framework, we conduct secondary training on health query medical texts to obtain dynamic text character vector representations. In the downstream task, we remove the original sentence-order prediction (SOP) [10] task from ALBERT and use the cross-entropy mechanism and sigmoid activation function to predict text labels. We construct a new task for text multi-label classification and perform supervised training through the backpropagation mechanism. For the medical text classification task based on transfer learning and ensemble learning, this paper uses the ALBERT pre-trained language model for transfer training and further extends this basic model, providing two variants of the model: the TLCM model and the TCLA model.

3. Method

3.1. TLCM Model Structure

The Bi-LSTM-CNN ensemble learning framework is constructed based on the RCNN hybrid neural network model framework. It introduces a CNN structure's filter behind multiple Bi-LSTM model structures to further convolve and pool the output vectors, forming a Bi-LSTM-CNN ensemble network module. Since the Bi-LSTM model takes into account the text's word order and semantic information, it has a strong modeling capability for texts with context-dependent relationships and is suitable for various specific task model constructions. It is currently a commonly used text semantic encoding model. This paper proposes a text classification model based on the Transformer structure and Bi-LSTM-CNN, fully considering the semantic information of the current character in the context, thereby learning the global structural information of the text. The Transformer-Bi-LSTM-CNN model framework is shown in Figure 2.

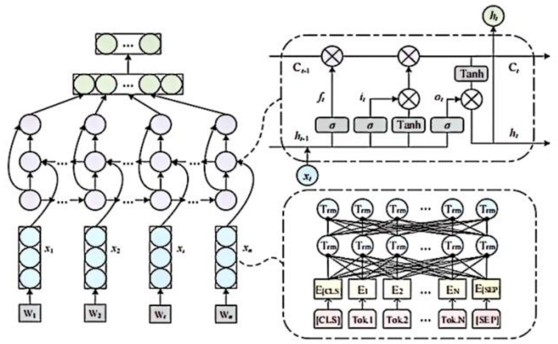

Figure 2. Bi-LSTM-CNN Ensemble Learning Framework

The Transformer-Bi-LSTM-CNN ensemble learning framework introduces the Transformer network structure from the traditional ALBERT model. The basic starting point of the model design is that the multi-layer bidirectional Transformer structure can achieve dynamic word vector representation of the text. In the downstream task, the Bi-LSTM model adopts a bidirectional sequence structure design, which can more deeply extract the context-dependent relationship information of the text sequence, achieving complex relationship modeling of sequence structure information. At the same time, the convolutional structure and pooling structure of the CNN model are used to further extract the semantic information features obtained by the Bi-LSTM model, more efficiently retaining local semantic information to construct a text label classifier.

Based on the Transformer-structured Bi-LSTM-CNN ensemble learning network framework, this paper introduces the version of the ALBERT pre-trained language model for transfer learning. It uses the original multi-layer bidirectional Transformer encoder within the ALBERT model for dynamic word vector representation of medical texts. On this basis, it constructs the Bi-LSTM-CNN ensemble learning module using the Bi-LSTM bidirectional long short-term memory network and the CNN convolutional neural network for medical text classification, achieving deep semantic information feature extraction of health query description texts. The output layer of the downstream subtask uses cross-entropy loss and sigmoid activation functions for multi-label classification training of texts, thus proposing a medical text multi-label classification algorithm based on Transformer, LSTM, and CNN singlechannel ensemble learning (Trans-LSTM-CNN-Multi, TLCM). The algorithm process is shown in Figure 3.

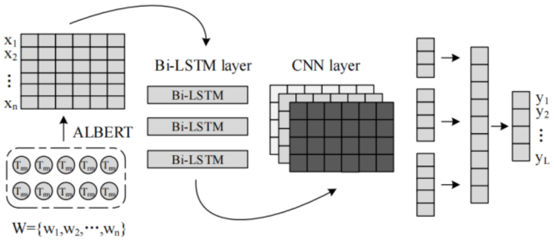

Figure 3. TLCM Model Framework

The TLCM medical text classification model introduces pre-trained ALBERT model for transfer training, which can enhance the model's semantic understanding of health query description texts. At the same time, it uses the Bi-LSTM bidirectional long short-term memory network to further strengthen the learning of text sequence context-dependent relationships, that is, global structural information features. Then, it uses the filter structure of the CNN convolutional neural network to further express the global structural information features obtained by the Bi-LSTM model into high-dimensional feature vectors, achieving a deeper level of text semantic extraction. To some extent, it can solve the problem of high-dimensional expression of medical text semantic information and insufficient knowledge utilization.

3.2. TLCA Model Structure

Medical texts are characterized by complex professional terminology and the common issue of polysemy, which affects the accuracy of model classification. CNN (Convolutional Neural Network) structures can effectively capture word-level information features of the text, while Bi-LSTM (Bidirectional Long Short-Term Memory) networks can learn the sequential structural information features of the text. In this paper, a multi-channel attention mechanism network layer is proposed for the integration of local feature information and global structural information. This layer obtains the output vectors of the CNN and Bi-LSTM structures, representing different dimensional information features, and performs feature fusion. By utilizing the Attention mechanism for integrated training of the information features output by both, a more comprehensive feature information expression for medical texts can be constructed.

The feature fusion operation between the CNN and Bi-LSTM structures inevitably leads to high vector dimensions, with issues such as sparse important information features and redundant feature vectors. If the model training process retains all information features, the model may struggle to focus on important information for learning. By introducing the Attention mechanism, the model can focus on important information during the training process, assign greater weight to important feature values, and downplay irrelevant information, thus training efficiently. This paper constructs an ensemble learning model based on the multi-channel attention mechanism, calculating the weight probability distribution of the output vectors expressed by the CNN and Bi-LSTM structures, as shown in Figure 4.

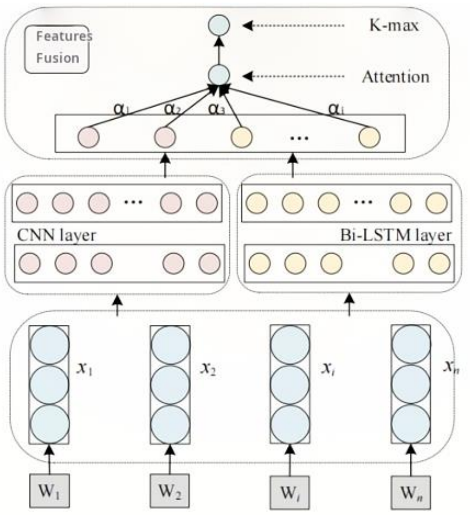

Figure 4. Multi-Channel Attention Mechanism

Similar to the TLCM model, the ensemble learning network framework based on the multi-channel attention mechanism includes a Transformer encoder, Bi-LSTM structure, CNN structure, and Attention mechanism. The model framework uses the original multi-layer bidirectional Transformer encoder within the ALBERT model for dynamic word vector representation of medical texts. On this basis, the Bi-LSTM structure and CNN structure further extract the output vectors of the upstream process Transformer structure, capturing the local semantic information features and global structural information features of the text, respectively. A Concatenation network layer and an Attention mechanism layer are added to fuse the vectors output by the CNN and Bi-LSTM structures, fully utilizing the important information features contained in the text. The output layer of the downstream subtask uses binary cross-entropy loss and sigmoid activation functions to construct a text multi-label classifier. Through model ensemble training, label prediction is achieved, thus proposing a multi-label classification algorithm for medical texts based on multi-channel attention mechanisms (Transformer-CNN-LSTM-Attention, TCLA). The overall framework of the algorithm is shown in Figure 5

The multi-label classification of texts based on the TCLA model framework integrates multiple neural network structures such as Transformer, CNN, LSTM, and Attention. The CNN structure can capture local lexical information features of the text, and the Bi-LSTM structure can extract contextual dependency relationship information features. The integrated application of the two can fully utilize the global and local semantic information features of the text, enhancing the model's semantic understanding of medical texts. The fusion of the output vectors of the CNN and Bi-LSTM structures inevitably leads to redundant feature vectors. The Attention mechanism focuses on important information features and downplays irrelevant information, which can reduce computational resource consumption to some extent, while also more fully and comprehensively utilizing important information features from different dimensions.

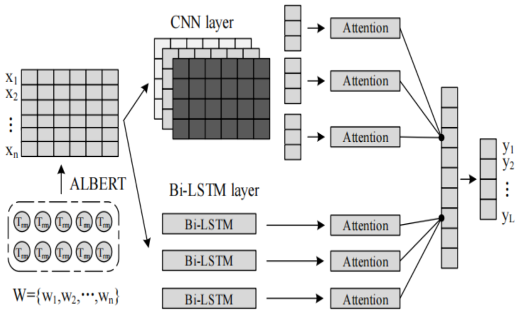

Figure 5. TCLA Model Framework

4. Experiment

4.1. Dataset

To assess the efficacy of the model presented in this paper, the study employed a public dataset comprising 5,000 online medical and health queries. This dataset originates from a text classification competition organized by a reputable medical association's informatics branch. Each sample within the health query dataset has been meticulously annotated by human experts with multiple medical categories, featuring an indeterminate number of labels that can range from zero to six. This characteristic categorizes it as a quintessential multi-label classification task dataset.

4.2.Experiments

This paper employs a version of the ALBERT pre-trained language model for transfer learning, and by fine-tuning the model, modifying the embedding network structure, and designing downstream classification subtasks, it proposes two multi-label medical text classification algorithms based on transfer learning and ensemble learning: TLCM and TCLA. To evaluate the performance of the TLCM and TCLA text classification algorithms based on transfer learning and ensemble learning, two ablation experimental models were also designed: Transformer-CNN and Transformer-Bi-LSTM.

In the aforementioned models, all four experimental groups used the original multi-layer bidirectional Transformer encoding structure within the ALBERT model for embedding and connected CNN structures and Bi-LSTM structures downstream of the Transformer structure for model ensemble training.

Table 1: Experiment Results

Model	Precision	Recall	Micro_F1
Transformer-CNN	89.50	89.69	89.60
Transformer-Bi-LSTM	88.51	88.61	88.51
TLCM（ours）	90.78	90.88	90.88
TCLA（ours）	90.59	90.68	90.68

To compare the performance differences among the models, the parameters involved in the four experimental groups remained consistent with the previous text, and each group used the same test and validation set data. The experimental results are shown in Table 1.

In the classification of health queries with small samples using the Transformer structure, the Transformer-CNN model achieved prediction accuracies of 89.5%, 86.69%, and 89.60% in Precision, Recall, and Micro-F1, respectively. The Transformer-Bi-LSTM model achieved prediction accuracies of 88.51%, 88.61%, and 88.51%, respectively. By optimizing the network structure and adjusting model hyperparameters, the TLCM model achieved the highest prediction accuracies of 90.78%, 90.88%, and 90.88%, obtaining the best multi-label classification effect. The TCLA model achieved prediction accuracies of 90.59%, 90.68%, and 90.68%, also demonstrating good classification performance. In addition, multiple experimental results indicate that the overall model multi-label classification accuracy, based on the fusion and ensemble training of the multi-layer bidirectional Transformer structure, CNN structure, and Bi-LSTM structure within the ALBERT model, is stable at around 91%, with a stable improvement of more than 1 percentage point compared to the prediction of a single structure of CNN and Bi-LSTM, further validating the feasibility of the text multi-label classification algorithm based on transfer learning and ensemble learning proposed in this paper.

5. Conclusion

This study presents a novel approach to medical text classification by leveraging the ALBERT pre-trained language model to address the complexities inherent in health query data, such as medical jargon, abbreviations, and linguistic challenges like synonyms, antonyms, and polysemy. To overcome the limitations of traditional classification methods—particularly in handling small datasets, hierarchical category labels, and high semantic similarities—this research proposes two innovative models: the TLCM and TCLA. These models integrate transfer learning and ensemble learning techniques, combining the strengths of CNN, Bi-LSTM, and Attention mechanisms to capture both local and global text features effectively. By fine-tuning ALBERT's multi-layer bidirectional Transformer structure and employing a dynamic word vector representation, the models were able to achieve superior performance, with approximately 91% in Precision, Recall, and Micro_F1. Furthermore, the models were optimized for multi-label classification, utilizing fully connected multi-layer perceptrons with cross-entropy and sigmoid activation functions to enhance the representation of context-specific information. The experimental results confirm that the proposed method significantly surpasses traditional text classification approaches, particularly in understanding complex medical semantics and improving classification accuracy. This study underscores the potential of pre-trained language models in advancing the field of medical text mining, offering a robust solution to the challenges associated with analyzing specialized health query texts. The combination of transfer learning and ensemble learning within this framework lays a strong foundation for further research and application in medical information retrieval and classification.

References

[1] Chang C H, Wang L, Yang C C, "Constructing cross-lingual consumer health vocabulary with Word-Embedding from comparable user generated content," Proceedings of the 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), pp. 275-284, 2024.

[2] Chiang C H, Huang S F, Lee H, "Pretrained language model embryology: The birth of ALBERT," arXiv preprint arXiv:2010.02480, 2020.

[3] Salehi A W, Khan S, Gupta G, et al., "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," Sustainability, vol. 15, no. 7, p. 5930, 2023.

[4] Yu Y, Yao Y, Liu Z, et al., "A Bi-LSTM approach for modelling movement uncertainty of crowdsourced human trajectories under complex urban environments," International Journal of Applied Earth Observation and Geoinformation, vol. 122, p. 103412, 2023.

[5] Mao A, Mohri M, Zhong Y, "Cross-entropy loss functions: Theoretical analysis and applications," Proceedings of the International Conference on Machine Learning, PMLR, pp. 23803-23828, 2023.

[6] Zhai X, Mustafa B, Kolesnikov A, et al., "Sigmoid loss for language image pre-training," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975-11986, 2023.

[7] Church K W, "Word2Vec," Natural Language Engineering, vol. 23, no. 1, pp. 155-162, 2017.

[8] Weiss K, Khoshgoftaar T M, Wang D D, "A survey of transfer learning," Journal of Big Data, vol. 3, p. 1-40, 2016.

[9] Parvin H, MirnabiBaboli M, Alinejad-Rokny H, "Proposing a classifier ensemble framework based on classifier selection and decision tree," Engineering Applications of Artificial Intelligence, vol. 37, pp. 34-42, 2015.

[10] Reczko M, "ELECTROLBERT: Combining Replaced Token Detection and Sentence Order Prediction," Proceedings of CLEF (Working Notes), pp. 335-340, 2022.