Original Research

DIGITAL HEALTH
Volume 9: 1–14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076231193213
journals.sagepub.com/home/dhj

Medical text classification based on the discriminative pre-training model and prompt-tuning

Yu Wang[1], Yuan Wang[2], Zhenwan Peng[1], Feifan Zhang[1], Luyao Zhou[1] and Fei Yang[1]

Abstract

Medical text classification, as a fundamental medical natural language processing task, aims to identify the categories to which a short medical text belongs. Current research has focused on performing the medical text classification task using a pre-training language model through fine-tuning. However, this paradigm introduces additional parameters when training extra classifiers. Recent studies have shown that the "prompt-tuning" paradigm induces better performance in many natural language processing tasks because it bridges the gap between pre-training goals and downstream tasks. The main idea of prompt-tuning is to transform binary or multi-classification tasks into mask prediction tasks by fully exploiting the features learned by pre-training language models. This study explores, for the first time, how to classify medical texts using a discriminative pre-training language model called ERNIE-Health through prompt-tuning. Specifically, we attempt to perform prompt-tuning based on the multi-token selection task, which is a pre-training task of ERNIE-Health. The raw text is wrapped into a new sequence with a template in which the category label is replaced by a [UNK] token. The model is then trained to calculate the probability distribution of the candidate categories. Our method is tested on the KUAKE-Question Intention Classification and CHiP-Clinical Trial Criterion datasets and obtains the accuracy values of 0.866 and 0.861. In addition, the loss values of our model decrease faster throughout the training period compared to the fine-tuning. The experimental results provide valuable insights to the community and suggest that prompt-tuning can be a promising approach to improve the performance of pre-training models in domain-specific tasks.

Keywords

Text classification, pre-training language model, prompt-tuning, ERNIE-Health, bidirectional encoder representations from transformers

Submission date: 13 December 2022; Acceptance date: 18 July 2023

Introduction

Medical texts such as medical literature and electronic medical records, contain valuable medical knowledge, including the symptoms, diagnosis, and medications of a particular disease.[1] Since learning such knowledge by human experts can be labor-intensive, natural language processing (NLP) has increasingly influenced the medical information research.[2,3] Medical text classification (MTC), as a fundamental medical NLP task, aims to identify the categories to which a short medical text belongs, such as disease stage, allergy intolerance, and organ status. The classification results affect the performance of downstream tasks, such as the detection of adverse medical events[4] or the construction of a clinical decision support system (CDSS).[5]

[1]School of Biomedical Engineering, Anhui Medical University, Hefei, China
[2]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China

Corresponding author:
Fei Yang, School of Biomedical Engineering, Anhui Medical University, Hefei, China.
Email: yangfei@ahmu.edu.cn

For the MTC task, previous research has established that the performance of machine learning[6] and statistical approaches[7] is highly dependent on the quality of feature engineering. In contrast, deep learning models, such as the convolutional neural network (CNN)[8] or the recurrent neural network (RNN),[9] can provide improved performance without additional manual feature selection. However, regardless of the approach being used, its essence is still to generate representations with abstract semantics and predict the categories based on them. Thus, considering that such models can only capture the semantic knowledge within training sets, they are unsatisfactory since the degree of improvement is still limited by the accessibility of labeled training sets. Therefore, how to enrich the semantic knowledge contained in the representations is the key to improving the performance of MTC.

Much of the current literature in NLP pays special attention to pre-training language models (PLMs), such as bidirectional encoder representations from transformers (BERT), enhanced language representation with informative entities (ERNIE), and ELECTRA.[10–12] Researchers have conducted extensive studies and suggested that PLMs have acquired rich prior semantic knowledge during the pre-training process, such as the masked language model (MLM) and next sentence prediction (NSP).[13,14] In the early stages of exploiting such knowledge for the MTC, researchers added additional classifiers on the top of PLMs and fine-tuned both of them using task-specific objective functions.[15,16] However, one major problem in such a "fine-tuning" paradigm is that additional parameters are introduced when tuning the extra classifiers. Therefore, how to bridge the gap between pre-training objectives and classification tasks has been considered as a critical factor in making a PLM more suitable for MTC.[17]

Recently, a number of studies have demonstrated that the "prompt-tuning" paradigm of PLMs induces better performance on a variety of NLP tasks.[18] The core concept of prompt-tuning is to use a pre-training task from the pre-training phase for a particular downstream task. For example, given that BERT uses the masked language model (MLM) during its pre-training phase, prompt-tuning methods based on the BERT would be built around the MLM task, which helps to reduce the difference between the pre-training task and the downstream task. A typical way to achieve such a paradigm is to wrap the input sentence in a natural language template and have a PLM perform the MLM. For instance, the sentence *"The patient has hypertension and diabetes."* shown in Figure 1 can be wrapped into the new input sequence with a template, and the probability of candidate words corresponding to the [MASK] token can be predicted by MLM. Prompt-tuning can be divided into two types: hard prompt-tuning and soft prompt-tuning. The former one is relatively simple, has a lower training cost, and closely resembles natural human language. However, a literature search revealed that no previous study has attempted to prove that prompt-tuning is also effective in classifying medical text.

In addition, current research on prompt-tuning has focused more on performing the downstream tasks based on the BERT series PLMs, ignoring that the discriminative PLMs, such as ELECTRA and ERNIE-Health[19] are strong alternatives. Inspired by the application of generative adversarial networks (GANs) in the field of computer vision (CV),[20] the discriminative PLMs aim to design a discriminator to distinguish whether a generator replaces a single token. It has been previously observed that the discriminative PLMs perform better in a variety of NLP tasks,[21] but their properties in the MTC remain unexplored.

The present research explores, for the first time, how to classify medical texts leveraging a discriminative PLM in a prompt-tuning paradigm. This study provides valuable insights to the community and suggests that prompt-tuning, can be a promising approach to improve the performance of pre-training models in domain-specific tasks. Moreover, this study mainly focuses on the fixed or hard prompt-tuning, as these methods are more concise. The main contributions of this study can be summarized as follows:

1. Firstly, this study provides new insights into MTC using the ERNIE-Health, a discriminative PLM consisting of a generator and a discriminator. ERNIE-Health is specifically designed for the medical domain, which means that it may have a better understanding of medical concepts and generate the representations with more abstract prior semantic knowledge than BERT, a more general-purpose model. This could lead to better performance on tasks that require domain-specific knowledge, such as the MTC task.
2. Secondly, in contrast to the existing prompt-tuning approaches based on MLM, this study attempts to perform prompt-tuning based on the multi-token selection (MTS) task, which is a pre-training task of ERNIE-Health. The rationale for choosing the MTS task for prompt-tuning in our study is based on the core concept of prompt-tuning, which is to use a pre-training task from the pre-training phase to reduce the discrepancy between the pre-training and downstream tasks. Unlike other discriminative PLMs, the discriminator of ERNIE-Health performs the MTS task in addition to the replaced token detection (RTD) task, that is, when the generator replaces a token, the MTS aims to predict the original token from a candidate word set. Specifically, we wrap the raw text into the new input sequence with a template, and use the [UNK] to replace the [MASK], forcing ERNIE-Health to predict the word replaced by [UNK] and achieving the prompt-tuning. In this way, the gap between the pre-training goals and classification tasks is bridged, and the category of a medical text can be inferred from the

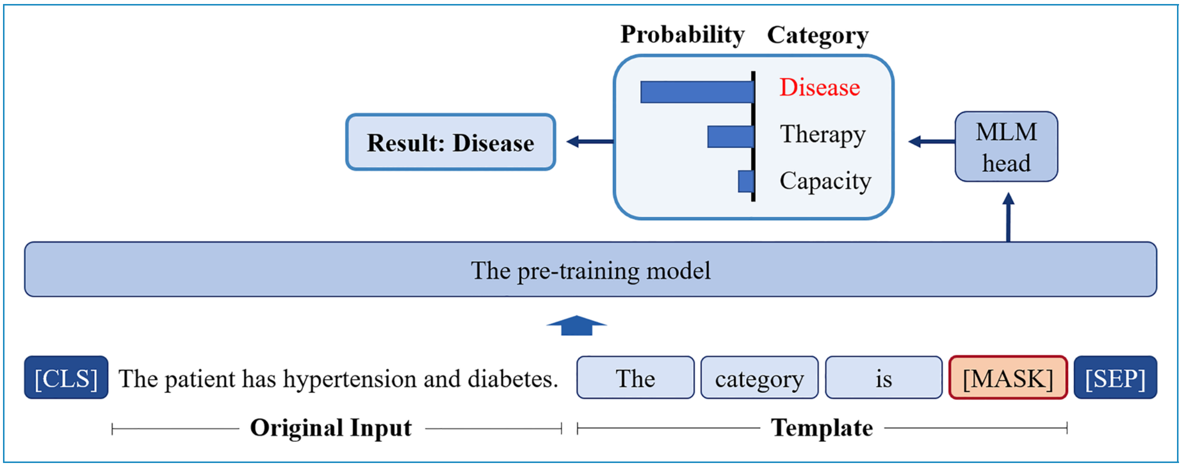

Figure 1. The illustration of prompt-tuning. The category "Disease" can be obtained by predicting the candidate word through masked language model (MLM).

predicted word without additional parameters or classifiers. Because prompt-tuning is a flexible method that allows for fine-grained control over the input to the model, the proposed method could be beneficial for the tasks that require specific or complex input formats, such as natural language inference or question answering.

3. Thirdly, we test our method on the KUAKE-Question Intention Classification (KUAKE-QIC) and CHiP-Clinical Trial Criterion (CHIP-CTC) datasets, which are sub-tasks in the Chinese Biomedical Language Understanding Evaluation (CBLUE).[22] The experimental results show that the accuracy values of our approach are between 0.866 and 0.861, which outperform the benchmark and previous approaches. In addition, the loss values of our model decrease faster throughout the training period compared to the method based on the fine-tuning paradigm.

In summary, our proposed prompt-tuning approach bridges the gap between pre-training and downstream tasks, leverages prior semantic knowledge, and leads to improved performance on specific tasks. We believe that this approach represents a promising direction for future research on MTC or other NLP tasks using PLMs in the medical domain.

The overall structure of this study takes the form of six sections. A brief review of the related work is presented in the "Related work" section. The "Methods" section deals with the methodology used in this study. The experimental results are presented in the "Experiments and results" section, while the discussion is provided in the "Discussion" section. Finally, the "Conclusions" section concludes this study with a summary.

Related work

Given that the method we propose falls under the theory of supervised learning, this section provides an overview of related work on MTC in supervised learning. Research on machine learning methods for MTC has a long history, with early examples including support vector machine (SVM) and hidden Markov model (HMM). For instance, Sarker and Gonzalez[6] used an SVM-based classifier to automatically detect adverse drug reactions from medical text, while Koopman et al.[23] trained SVM classifiers with detailed features to identify cancer-related causes. Kocbek et al.[24] presented a text-mining system to detect positive cancer admissions, while Yi and Beheshti[7] tested the application of HMM in medical text classification. However, the manual selection of appropriate features remains a key challenge, as it is time-consuming and labor-intensive. Therefore, the existing literature on MTC mainly focuses on identifying the categories based on deep learning and pre-training models.

Deep learning

Deep learning methods have become popular since 2012 with the improvement of computer hardware, especially the graphics processing unit (GPU). Over the past decade, most research in the field of MTC has emphasized the use of two standard deep learning models: the CNNs[8] and RNNs.[9]

Convolutional neural networks. The CNN, originally used in CV,[25] consists of convolutional kernels and pooling layers. In NLP, researchers have applied CNNs to generate word embeddings, capturing local features through pooling. Rios and Kavuluru[26] showed that CNNs outperformed previous approaches in biomedical text classification for assigning medical subject headings. Hughes et al.[27] applied CNNs to sentence classification in the Merck Manual dataset, the first study of its kind in medical texts. Nii et al.[28] used CNNs to classify nursing-care texts and extract essential parts for classification. Yao et al.[29] combined rule-based features with a knowledge-guided CNN model trained with word and entity embeddings from Unified Medical Language System.

Recurrent neural networks. RNNs are more suitable than CNNs for sequence modeling, which is crucial for text classification with sequential data. In medical text classification, the long short-term memory (LSTM) and gated recurrent unit (GRU), both RNN variants with gated units, are more widely used due to their better handling of gradient vanishing and explosion. Oleynik et al.[30] evaluated machine learning and LSTM for multi-label binary text classification. Venkataraman et al.[31] trained an RNN to automate clinical record assignment. Liang et al.[32] improved LSTM with a dual-channel mechanism for Chinese medical text classification. Bangyal et al.[33] applied RNN and LSTM to COVID-19 fake news detection, a medical text classification task.

Hybrid methods and attention. A number of studies have also proposed methods that take advantage of both CNNs and RNNs. Zhou et al.[34] used an integrated CNN–RNN framework to extract semantic and sequential features from patient queries. Li et al.[35] proposed a three-stage hybrid method using BiLSTM and a regular expression-based classifier for medical text classification. Ibrahim et al.[36] used a CNN and LSTM in a deep generic learning-based hybrid multi-label classification method. Shin et al.[37] combined the attention mechanism and CNN to classify radiology head CT reports and generated a heat map of attended terms.

As can be seen, the core of these deep learning approaches is to generate representations with semantic knowledge using CNNs or RNNs. However, the semantic knowledge can only be abstracted from the training sets. In other words, the performance of these methods is still limited by the inaccessible labeled training sets.

Pre-training language models

Much of the current literature on medical text classification pays special attention to pre-training models. Inspired by the ImageNet[38] in the field of Computer Version, researchers have proposed pre-training language models, such as BERT, ERNIE, ELECTRA, and ERNIE-Health for NLP. All PLMs adopt a hierarchical architecture consisting of several layers of transformer blocks as shown in Figure 2. Each transformer block contains a self-attention mechanism[39] that allows the model to pay attention to the different parts of the input sequence and to capture long-range dependencies. These models are pre-trained on large-scale text corpora using various pre-training tasks to learn contextual representations of words and sentences. For example, BERT uses the MLM and the NSP task, while discriminative PLMs such as ELECTRA and ERNIE-Health have an additional discriminator to determine whether a token is “Original” or “Replaced,” that is, the Replaced Token Detection (RTD) task, which allows them to generate representations with more abstract semantics. In addition, ERNIE-Health uses a multi-task learning approach with tasks such as RTD and MTS to capture domain-specific knowledge and semantics relevant to the medical domain. As shown in Figure 3, ERNIE-Health consists of a generator and a discriminator. The generator is used to perform an MLM task that infers what the masked word is in the original sentence. The discriminator determines sequentially whether each token has been replaced, that is, the same RTD task as in ELECTRA. However, based on the RTD results, ERNIE-Health also selects the original word of the replaced word from the candidate word set, that is, the MTS task.

There are two paradigms for transferring prior knowledge to a specific downstream task: fine-tuning and prompt-tuning. For the first one, extra classifiers are added on the top of PLMs, and both are tuned using task-specific objective functions. Almost all current work on MTC based on the PLMs follows this paradigm. As an example, Qasim et al.[15] detected the fake news of COVID-19 by tuning BERT, RoBERTa, and XLNet. Guo et al.[16] evaluated six types of PLMs on social media health-related text classification tasks. An elaborate experiment conducted by Lu et al.[40] verified that BERT performed best in all scenarios for classifying the presence or the absence of 16 diseases from patient discharge summaries. Peng et al.[41] found that the BERT model pre-trained on PubMed abstracts and MIMIC-III clinical notes performed best when tested on the biomedical language understanding evaluation (BLUE) benchmark. However, Gao et al.[42] showed that BERT often fails to outperform these simpler benchmarks when classifying MIMIC-III discharge summaries and cancer pathology reports, limiting the application of transformers for document classification on long clinical texts.

The disadvantage of fine-tuning is that it introduces additional parameters and tunes extra classifiers. Recently, a number of studies have demonstrated that prompt-tuning induces better performance for NLP tasks. In such a paradigm, the input sentence is wrapped into a natural language template in order to bridge the gap between pre-training goals and downstream tasks. For example, Jiang et al.[43] proposed a knowledgeable prompt-tuning method for the fake news detection task. Li et al.[44] proposed a knowledge-injected prompt-tuning model to detect events from the text by identifying and classifying event triggers. To address the problem that the number of labeled texts is small due to the lack of specialized expertise, a prompt-tuning method for

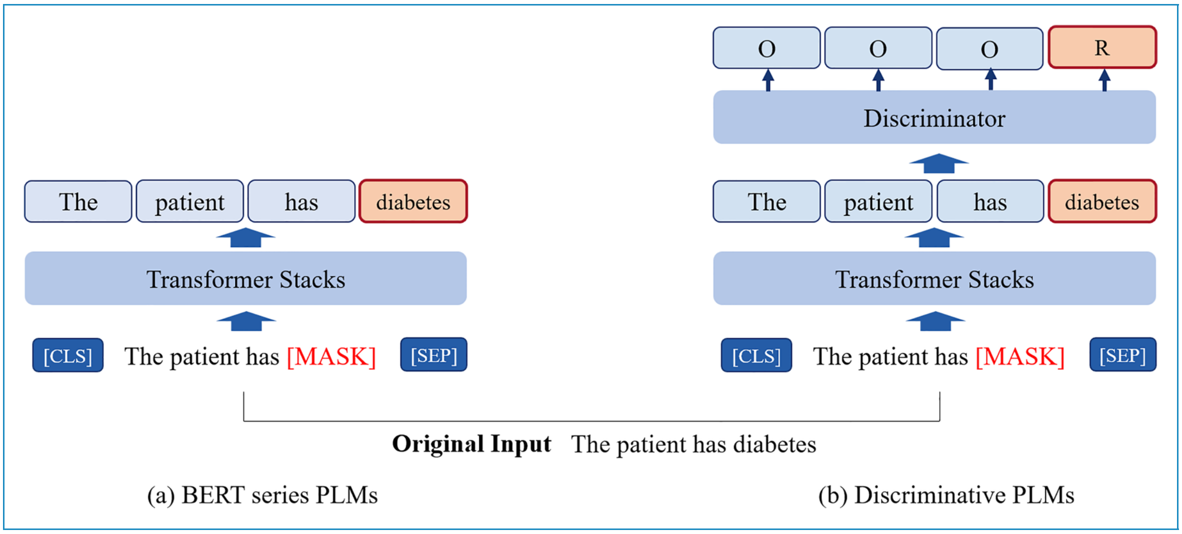

Figure 2. The difference between BERT series PLMs and discriminative PLMs. BERT: bidirectional encoder representations from transformers; PLMs: pre-training language models.

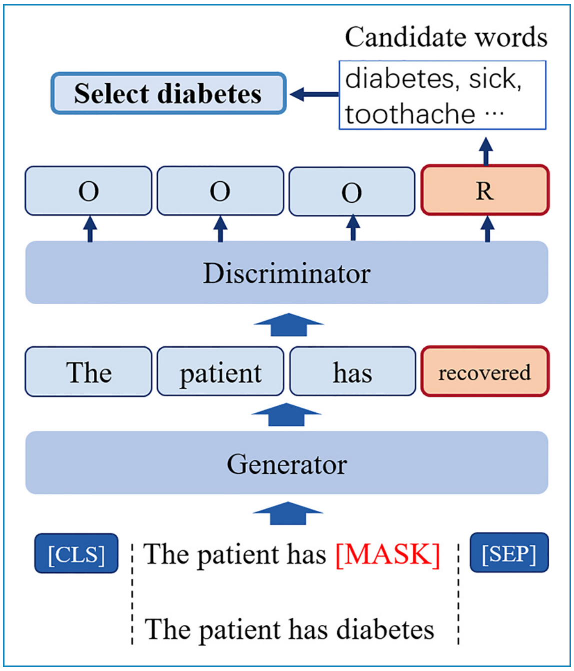

Figure 3. The overview of enhanced language representation with informative entities (ERNIE)-Health.

multi-label text classification was introduced by Wei et al.[45] However, a literature search revealed that no previous study has attempted to prove that prompt-tuning is also effective in classifying the medical text.

Methods

This section first defines the MTC task, and then the proposed model is introduced in detail.

Problem definition

For the MTC task, the input is a medical related sentence, while the output is the category to which the sentence belongs. Give a dataset $D = \{X, Y\}$, where X denotes the set of training sentences, and Y denotes the set of labels corresponding to the training set. For any $x_i \in X$, and $y_j \in Y$, if there exists a "belong-to" relation between x_i and y_j, then it is denoted as $d_{i,j} \in D$. The mapping function $f(x_i)y_j$ is calculated through the dataset D. For another dataset $D' = \{X', Y'\}$ with the same distribution of labels as D, there exists $d'_{i,j} = \{x'_i, y'_j\}$ and $d'_{i,j} \in D'$. The $\hat{y}'_j = f(x'_i)$ should be as close as possible to the true value y'_j. Therefore, the key of the MTC task is to find an optimal model to determine the mapping function $f(x_i)y_j$.

Model architecture

The proposed model is shown in Figure 4, and the selected pre-training model is the discriminator of ERNIE-Health, as it is pre-trained on medical corpora. In contrast to the most common paradigm fine-tuning, we utilize prompt-tuning to formalize the classification task into a multi-token selection problem, which is identical to the pre-training process.

(1) Input part

Give a text sequence $X = \{x_0, x_1, x_2, x_n\}$, a mapping function $f(\cdot)$ is used to wrap this sentence into X_p with a *template*, which is a piece of natural language text. As an example, the raw text X in Figure 4 is *"What is the best therapy plan for diabetes."*, and the intent of this statement should be classified as "treatment." Then, we wrap X into X_p according to the following mapping function $f(\cdot)$:

$$f(X) = \texttt{[CLS]}\ A\ \texttt{[UNK]}\ intent:\ X\ \texttt{[SEP]}. \qquad (1)$$

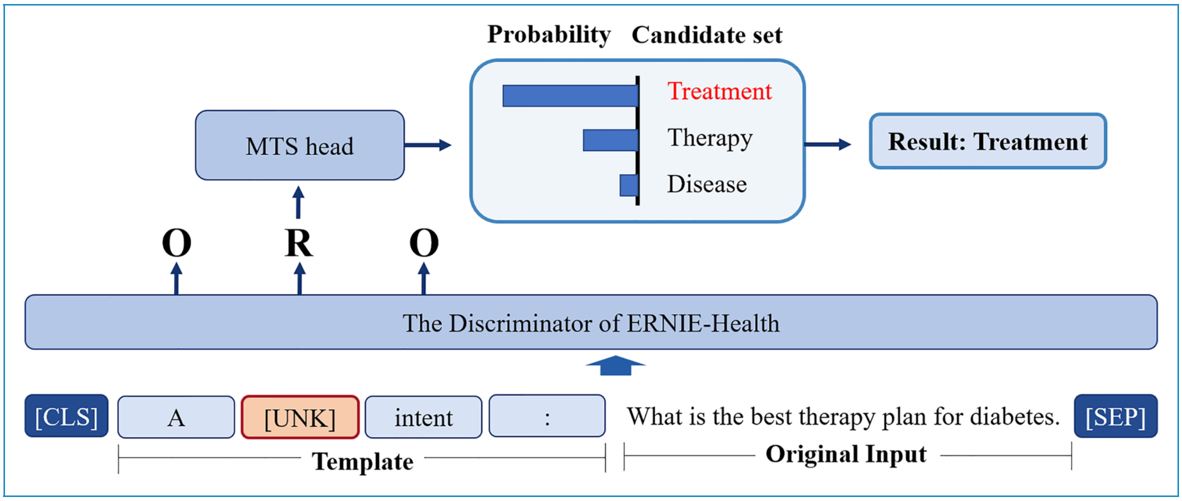

Figure 4. The architecture of the proposed model.

We use the [UNK] token as a placeholder for the category label in the language template. This token will be marked as "replaced" in the output of the discriminator manually, and the true word will be predicted during the MTS process. In other words, in this case, the model is trained to predict "treatment" at the position corresponding to the [UNK] token in X_p.

(2) Output part

Then, X_p is fed into the discriminator of ERNIE-Health to obtain the representations $H = \{h_{[\text{CLS}]}, h_1, h_2, h_{\text{SEP}}\}$ of input tokens. As mentioned above, ERNIE-Health conducts two kinds of token-level pre-training tasks on the part of the discriminator: RTD and MTS. The RTD aims to detect whether a token is original or replaced, and once the token is replaced, the MTS aims to select the original word from a candidate set S_c. In our model, the [UNK] token is marked as "replaced" manually, and an MTS task is conducted to predict the original word. Expressly, the candidate set S_c consists of category words, and the model will calculate the probability $P([\text{UNK}] = w|X_p, S_c)$ of each word w in the candidate S_c according to the following equation:

$$P([\text{UNK}] = w) = \frac{\exp\,[e(w)^T h_{[\text{UNK}]}]}{\sum_{w' \in S_c} \exp\,[e(w')^T h_{[\text{UNK}]}]}, \qquad (2)$$

where $e(\ldots)$ is the embedding lookup operation. Given that the MTS operation is essentially a multi-classification problem, the loss function is identical to the cross-entropy error as the following equation:

$$loss = -\sum_{n=1}^{N} p(w^n) \log\,[q(w^n)], \qquad (3)$$

where $p(w^n)$ denotes the probability distribution of the correct word (i.e. the original word) to replace [UNK] and $q(w^n)$ denotes the probability distribution of the predicted word. The training goal is minimizing the loss function above.

Experiments and results

This section introduces the dataset used in the experiment and shows the performance of the proposed model. The goal of this experiment is to validate the effectiveness of the proposed method by evaluating the classification accuracy of all models listed in Tables 1 to 4 for the MTC task, using the "KUAKE-QIC" and "CHIP-CTC" datasets. Since previous models were based on fine-tuning, we first tested the performance of fine-tuning with ERNIE-Health, and then evaluated its performance with prompt-tuning. The software environment for this experiment is the Paddlepaddle[1] , an end-to-end open-source deep learning platform developed by Baidu. In our experiments, we used a server equipped with an 8-core CPU and a NVIDIA Tesla V100 GPU, and 16 GB of RAM.

Dataset

The datasets used for our experiment are the "KUAKE-QIC" and "CHIP-CTC," which are sub-tasks of the CBLUE[22] dataset.[2] The KUAKE-QIC contains short Chinese texts related to patient inquiries and has proven to be a valuable resource for the development and evaluation of natural language processing techniques. Each sample in the dataset is assigned a label indicating one of the following intentions: "Diagnosis," "Cause," "Method," "Advice," "Metric explain," "Disease express," "Result," "Attention," "Effect," or "Price." If the intent of a sample does not fall into one of these categories, it is labeled as"Other." The dataset provides a diverse range of intentions that a patient may have when seeking medical advice. This diversity allows for a more comprehensive evaluation of the performance about the proposed method on various patient inquiries. The train set contains a total of 6931 samples, and the distribution of the text labels is presented in Table 2. The

Table 1. The distribution of labels in the CHiP-clinical trial criterion (CHIP-CTC) dataset.

Labels	Train set	Test set
Disease	5127	1693
Symptom	154	52
Sign	286	91
Pregnancy-related activity	1026	342
Neoplasm status	131	49
Non-Neoplasm disease stage	103	34
Allergy intolerance	668	227
Organ or tissue status	358	120
Life expectancy	166	56
Oral-related	51	18
Pharmaceutical substance or drug	877	298
Therapy or surgery	1504	487
Device	129	39
Nursing	22	12
Diagnostic	1233	412
Laboratory examinations	1142	374
Risk assessment	708	233
Receptor status	28	10
Age	917	304
Special patient characteristic	104	33
Literacy	52	18
Gender	30	10
Education	19	8
Address	31	11
Ethnicity	13	5
Consent	1319	448
Enrollment in other studies	514	172
Researcher decision	464	152
Capacity	168	50
Ethical audit	12	11
Compliance with protocol	370	130
Addictive behavior	272	90
Bed time	14	11
Exercise	21	7
Diet	61	20
Alcohol consumer	17	6
Sexual related	17	13
Smoking status	54	19
Blood donation	31	9
Encounter	66	28
Disabilities	17	8
Healthy	39	12
Data accessible	71	24
Multiple	4556	1536
Total	22,962	7682

distribution of labels in the dataset indicates that the most common intention is “Query for method,” which accounts for 25% of the total samples.

The CHIP-CTC dataset contains 44 pre-defined semantic categories for filtering criteria and a set of descriptive sentences for Chinese clinical trial screening criteria. The goal of this task is to determine the specific category for each screening criterion. It can be seen from Table 1 that the training sample size of CHIP-CTC is three times larger than that of KUAKE-QIC, and the number of categories is four times larger.

Hyper-parameters

The hyper-parameters involved in this experiment are listed in Table 3. We used Adam as the optimizer and trained our

Table 2. The distribution of labels in the KUAKE-question intention classification (KUAKE-QIC) dataset.

Labels	Train set	Test set
Diagnosis	877	288
Cause	153	29
Method	1750	676
Advice	371	134
Metric explain	137	32
Disease express	594	158
Result	235	45
Attention	650	120
Effect	370	28
Price	177	50
Other	1617	395
Total	6931	1955

Table 3. Hyper-parameters.

Hyper-parameters	Values	
	KUAKE-QIC	CHIP-CTC
Epoch	3	5
Learning rate	6×10^{-5}	5×10^{-5}
Input max length	64	128
Batch size	16	32

KUAKE-QIC: KUAKE-question intention classification; CHIP-CTC: CHiP-clinical trial criterion.

model for three epochs with a batch size of 16 with KUAKE-QIC, and for five epochs with a batch size of 32 with CHIP-CTC.

Evaluation metrics

We introduce the accuracy to evaluate the performance of the models listed in Table 4. This metric is calculated according to the following formulation where n denotes the number of samples that predict correctly, and N denotes the total number of samples.

$$Accuracy = \frac{n}{N} \tag{4}$$

Results of different models

First, we tested the performance of some popular PLMs on the KUAKE-QIC dataset, and the experimental results are shown in Table 4, where the BERT-base is the baseline model. As mentioned before, BERT[10] is a PLM that encodes the representations with prior semantic and syntactic knowledge through pre-training tasks MLM and NSP. It is selected as the baseline model in our experiment, and the baseline accuracy is 0.843 with fine-tuning. BERT-wwm-ext-base is an extended version of BERT-base. This PLM is retrained based on the initial checkpoint of the BERT-base and performs the pre-training task on an extended corpus with Whole Word Mask (WWM). Given that the WWM is more suitable for a Chinese MTC because there is no separator between Chinese words, BERT-wwm-ext-base improves the accuracy to 0.845 with fine-tuning. Furthermore, the base and large versions of MacBERT[47] are all extended PLMs of BERT-base. The difference between these two kinds of versions lies in the number of whole parameters and the dimensional size of the hidden layers. The MacBERT optimized the WWM of BERT-wwm-ext-base with N-gram masking strategies, and the two types of versions of MacBERT gets the accuracy of 0.849 and 0.827 with fine-tuning, respectively. Moreover, the RoBERT-large[48] utilizes a random masking strategy and obtains the representations with richer semantics compared to the static masking strategy of BERT. The RoBERTa-wwm-ext-large improved RoBERT-large using WWM, and this PLM obtained the highest accuracy of 0.853 among the PLMs listed in Table 4 except ERNIE-Health.

Given that the ERNIE-Health is pre-trained on medical corpora, it gets an accuracy of 0.844 when fine-tuning, slightly higher than the baseline. The approach we proposed is to classify the medical texts with prompt-tuning using ERNIE-Health, and this model obtains an accuracy of 0.866, which is higher than the result of RoBERTa-wwm-ext-large. Moreover, the architecture of the latter one is more complex than it of our model. The number of transformer block layers and hidden layer sizes are 12 and 768 in our model, while these two indices are 16 and 1024 in RoBERTa-wwm-ext-large, respectively. That means the RoBERTa-wwm-ext-large may need more training time but obtain a lower accuracy value than our model.

Second, we also tested the performance of the PLMs on the CHIP-CTC dataset, and the experimental results are shown in Table 4, where the BERT-base is also the baseline model. The baseline accuracy is 0.854 with fine-tuning. BERT-wwm-ext-base, as an extended version of BERT-base, improves the accuracy to 0.856 with fine-tuning. Furthermore, for the base-version and large-version of MacBERT,[47] the two types of versions of MacBERT get the accuracy of 0.854 and 0.855 with fine-tuning,

Table 4. The results of different models on the test set.

Models	KUAKE-QIC			CHIP-CTC		
	Fine-tuning	Prompt-tuning	ΔA	Fine-tuning	Prompt-tuning	ΔA
BERT-base (baseline)[10]	0.843	0.848	+0.05	0.854	0.855	+0.01
BERT-wwm-ext-base[46]	0.845	0.846	+0.01	0.856	0.858	+0.02
MacBERT-base[47]	0.849	0.851	+0.02	0.855	0.854	−0.01
MacBERT-large[47]	0.827	0.846	+0.19	0.856	0.855	−0.01
RoBERTa-wwm-ext-large[48]	0.853	0.858	+0.05	0.857	0.858	+0.01
ERNIE-Health	0.844	0.866	+0.22	0.857	0.861	+0.04
Average	—	—	+0.09	—	—	+0.01

KUAKE-QIC: KUAKE-question intention classification; CHIP-CTC: CHiP-clinical trial criterion; BERT: bidirectional encoder representations from transformers; ERNIE: enhanced language representation with informative entities.

Table 5. The results of *T*-tests. The hypothesis of the *T*-test is that there is no significant difference in the means between the two groups of random experiments.

Models	*p*-values	
	KUAKE-QIC	CHIP-CTC
BERT-base (baseline)	1.967×10^{-4}	0.06125
ERNIE-Health	1.465×10^{-10}	4.078×10^{-7}

KUAKE-QIC: KUAKE-question intention classification; CHIP-CTC: CHiP-clinical trial criterion; BERT: bidirectional encoder representations from transformers; ERNIE: enhanced language representation with informative entities.

respectively. Furthermore, the RoBERTa-wwm-ext-large, which is the improved version of the RoBERT-large[48] obtains the highest accuracy of 0.858 among the PLMs listed in Table 4 except ERNIE-Health.

The ERNIE-Health obtains an accuracy of 0.857 when fine-tuning, which is slightly higher than the baseline. The approach we proposed achieves an accuracy of 0.861, which is higher than the result of RoBERTa-wwm-ext-large, but the improvement is modest. Similarly, given that the architecture of the latter one is more complex than that of our model, the RoBERTa-wwm-ext-large may need more training time but obtain a lower accuracy value than our model on the CHIP-CTC dataset, too.

Results of fine-tuning and prompt-tuning

Moreover, the performance of two types of paradigms: fine-tuning and prompt-tuning based on the PLMs listed in Table 4 are tested on both the KUAKE-QIC and CHIP-CTC datasets. The experimental results are shown in Table 4, where all the models based on the prompt-tuning outperform those based on the fine-tuning when testing on KUAKE-QIC, increasing the accuracy with an average value of 0.09. However, the performance of the proposed method on CHIP-CTC is not significant when compared with KUAKE-QIC. Although the prompt-tuning based on ERNIE-Health can still improve the accuracy by 0.04, the average accuracy is only increased by 0.01, and even the accuracy of the MacBERT series models is decreased.

Based on the findings presented in Table 4, it is evident that the performance improvement of prompt-tuning is relatively small for some PLMs. Consequently, we conduct additional experiments to validate the statistical significance of this accuracy improvement. Specifically, we conduct random experiments based on fine-tuning and prompt-tuning, respectively, and the seeds are randomly assigned. Subsequently, a paired *T*-test is performed to compare the accuracy values obtained from two random experimental groups. Given the limited sample size (the number of each random experiments group is 10), a Shapiro–Wilk test is performed to assess the normality of the data prior to conducting the *T*-test. The *p*-values of the *T*-tests conducted on the fine-tuning and prompt-tuning, utilizing both BERT-base and ERNIE-Health models, are presented in Table 5. Notably, for the KUAKE-QIC dataset, all *p*-values are found to be less than the significance level of 0.05, indicating the statistical significance for the improvement of prompt-tuning. For the CHIP-CTC dataset, when we utilize ERNIE-Health, a significant difference is observed between the fine-tuning and prompt-tuning paradigms, with a *p*-value from the

T-test below the significance level of 0.05. However, when we use BERT, the *p*-value is slightly larger than 0.05. The details of the random experiments are presented in the Appendix.

In addition, we also record the loss-values at each training batch for fine-tuning and prompt-tuning based on the ERNIE-Health. As shown in Figure 5(a), the red line represents the loss-values of fine-tuning and the blue line represents those of prompt-tuning. It is clear from this figure that the prompt-tuning curve decreases more rapidly than the fine-tuning curve, especially in the early stages of training. However, since the loss-values are oscillating and decreasing, we also provide the smooth version of the loss curves to compare the two types of paradigms more clearly, which can be found in Figure 5(b) and (c). The loss curves, as shown in Figure 5(b), show more obvious that the loss-values of prompt-tuning fall faster in the early stages. Furthermore, Figure 5(c) illustrates the results when the loss-values are taken logarithmically and shows that the loss-values of prompt-tuning also decrease faster at the later stage.

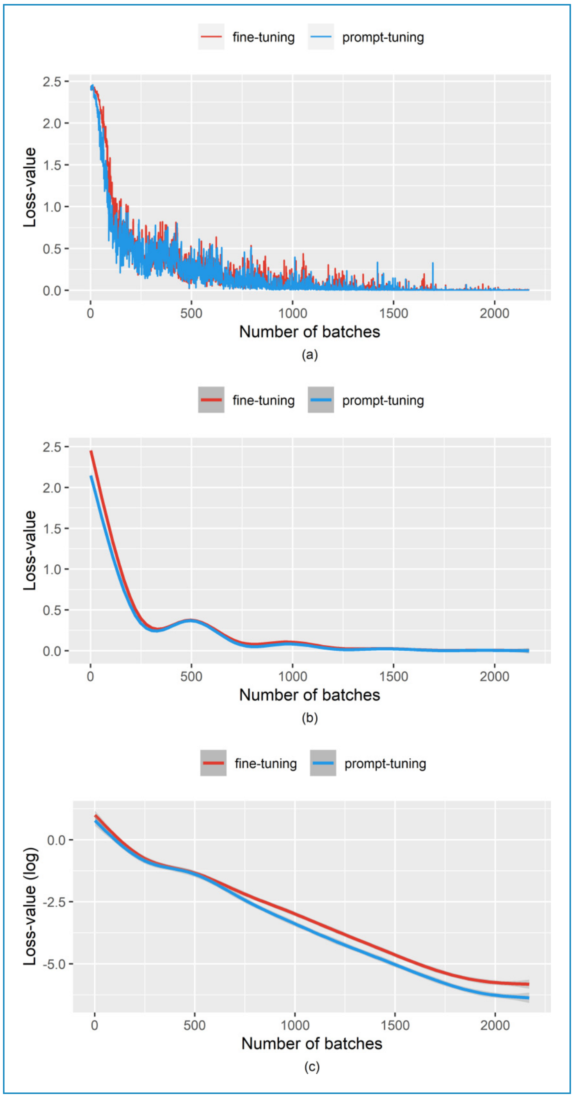

Figure 5. (a) The loss curves of fine-tuning and prompt-tuning, (b) the smoothing curves of loss curves, and (c) the smoothing curves of loss curves under logarithm.

Results of different templates

Considering that the templates used in prompt-tuning may affect the performance of text classification, we conducted experiments to test the impact of different templates on the accuracy of text classification, as shown in Table 6. We differentiate these templates based on whether they are closer to the natural language used by humans. From Table 6, we can see that the closer the template is to the natural language used by humans, the higher the accuracy obtained by prompt-tuning.

Discussion

We tested our method on the KUAKE-QIC and CHIP-CTC datasets and compared the results with those obtained by some popular PLMs. It can be seen from Table 4 that our method, for example, ERNIE-Health with prompt-tuning, achieves the highest accuracy on both datasets. It can increase the accuracy by 0.08 and 0.03 compared to the RoBERTa-wwm-ext-large. In addition, the total parameters and the hidden layers of our method are smaller than those of RoBERTa-wwm-ext-large. Therefore, our model spends less time on training and is less likely to overfit, which makes it more competitive.

The results presented in Table 4 provide strong evidence that prompt-tuning is an effective method for improving the performance of PLMs. Essentially, the prompt-tuning paradigm still performs a classification task, that is, calculating the probability of the output word corresponding to the [MASK] token in the template. The process still involves computing each word in the entire word list and updating the weights of the PLM which may still be downstream task oriented. However, unlike the fine-tuning paradigm

Table 6. The results of different templates.

Templates	Accuracy values
The [MASK]: [RAW]	0.8556
A [MASK] intent: [RAW]	0.8616
The intent is [MASK]: [RAW]	0.8621
The intent of [RAW] is [MASK]	**0.8661**

Table 7. The results of random experiments on KUAKE-QIC.

Models	Paradigms	Accuracy values	*p*-values (Shapiro–Wilk)	*p*-values (*T*-test)
BERT-base	Fine-tuning	0.842131 0.843151 0.839061 0.843671 0.842641 0.843151 0.846221 0.842641 0.844181 0.843151	0.1716	1.967×10^{-4}
	Prompt-tuning	0.848207 0.846157 0.848207 0.848717 0.851787 0.843087 0.849737 0.845137 0.849737 0.849227	0.6121	
ERNIE-Health	Fine-tuning	0.842719 0.844769 0.844259 0.845279 0.844769 0.845279 0.843749 0.841699 0.844259 0.843229	0.3743	1.465×10^{-10}
	Prompt-tuning	0.863696 0.867796 0.867276 0.865746 0.865236 0.862676 0.864716 0.866256 0.869326 0.867276	0.9924	

KUAKE-QIC: KUAKE-question intention classification; BERT: bidirectional encoder representations from transformers; ERNIE: enhanced language representation with informative entities.

Table 8. The results of random experiments on CHIP-CTC.

Models	Paradigms	Accuracy values	*p*-values (Shapiro–Wilk)	*p*-values (*T*-test)
BERT-base	Fine-tuning	0.855690 0.855040 0.853610 0.852180 0.852180 0.855560 0.853610 0.854000 0.854780 0.853350	0.4304	0.06125
	Prompt-tuning	0.858264 0.853974 0.854104 0.856834 0.852924 0.857744 0.852674 0.853584 0.854364 0.855534	0.2444	
ERNIE-Health	Fine-tuning	0.858248 0.857208 0.857988 0.856948 0.857468 0.855908 0.858248 0.855778 0.855518 0.856688	0.358	4.078e-07
	Prompt-tuning	0.860800 0.862230 0.861060 0.861580 0.861190 0.860150 0.860800 0.861580 0.860020 0.860600	0.8976	

CHIP-CTC: CHiP-clinical trial criterion; BERT: bidirectional encoder representations from transformers; ERNIE: enhanced language representation with informative entities.

that uses the [CLS] token, prompt-tuning constructs a template with the [MASK] token based on the raw input and then allows the PLM to predict the output. The prompt-tuning paradigm helps bridge the gap between pre-training and downstream tasks, resulting in improved performance on specific tasks. In contrast to the fine-tuning approach adopted by other PLMs listed in Table 4, where additional classifiers are added to the PLMs and both are fine-tuned using task-specific objective functions, our proposed model performs prompt-tuning leveraging the pre-training task: MTS. This approach can be thought of as a kind of MLM, where we wrap the raw text into a new input sequence with a template and train the model to predict the correct token to be replaced by [UNK] in the template. The reason we use [UNK] to replace [MASK] is to force the ERNIE-Health discriminator to select a word from the candidate word set to replace [UNK], thus achieving prompt-tuning.

Importantly, our model is trained based on the initial parameters that are the same as those of the pre-trained ERNIE-Health, allowing us to fully exploit the prior semantic knowledge encoded in ERNIE-Health. Moreover, ERNIE-Health is pre-trained on medical corpora, and its representations contain semantic knowledge relevant to the medical domain. Therefore, our proposed method is particularly well-suited for tasks in the medical domain. We note that the fine-tuning paradigm introduces additional parameters, and the training process starts from scratch, making it computationally expensive and time-consuming. In contrast, our prompt-tuning approach is more efficient and does not require additional parameters to be added to the model. However, we found that the performance improvement of our proposed method on the CHIP-CTC dataset is limited, and in some cases, it is even inferior to fine-tuning. We believe that the reason for this is related to the size of the training set. The number of training samples of CHIP-CTC is four times higher than that of KUAKE-QIC. At this point, the advantage of prompt-tuning in bridging the gap between pre-training tasks and specific downstream tasks may no longer be significant.

Furthermore, the experimental results presented in Table 4 show that the accuracy of ERNIE-Health is slightly higher than that of BERT-base when performing fine-tuning. This observation highlights the importance of pre-training on domain-specific corpora and leveraging prior semantic knowledge when training PLMs for specific downstream tasks.

In addition, we also show the loss curves for ERNIE-Health based on fine-tuning and prompt-tuning. As shown in Figure 5, the loss-values of prompt-tuning decrease faster than those of fine-tuning, especially in the early stages of training. Moreover, when plotting the smooth version of the loss curve and illustrating the logarithmical loss value, another important finding is that the loss-values of prompt-tuning are lower than those of fine-tuning not only in the early stage but also in the later stage. This result can be explained by the fact that our model (ERNIE-Health with prompt-tuning) is trained based on the initial parameters that are the same as those of a trained ERNIE-Health. Therefore, the model can converge faster considering that the fine-tuning paradigm requires additional classifiers to be trained from scratch.

We also conduct random experiments to validate the statistical significance of the accuracy improvement presented in Table 4. The *p*-values of four paired *T*-tests listed in Table 5 provide the evidence that for the KUAKE-QIC dataset, irrespective of whether BERT-base or ERNIE-Health is employed, the *p*-values derived from the *T*-test are below the significance level of 0.05. Thus, it can be concluded that prompt-tuning demonstrates significantly improvement relative to fine-tuning on this particular dataset. For the CHIP-CTC dataset, when ERNIE-Health is utilized, the *p*-value derived from the *T*-test falls below the significance level of 0.05. However, when BERT-base is employed, the *p*-value marginally surpasses 0.05. Consequently, given that the training sample size of the CHIP-CTC dataset is three times greater than that of the KUAKE-QIC dataset, it can be suggested that prompt-tuning may exhibit more favorable performance on datasets with smaller-scale training samples. This is consistent with the conclusions we obtained from Table 4.

Finally, we also tested the performance of different templates on prompt-tuning. As shown in Table 6, the closer the template is to natural language, the higher the accuracy obtained. This is again due to the nature of prompt-tuning. The closer the template is to natural language, the more likely the PLM has “seen” it during pre-training. Therefore, the probability of predicting the correct replacement word is higher.

Conclusions

This study provides new insights into classifying medical text using a discriminative PLM with prompt-tuning. The discriminative PLM selected for this study is ERNIE-Health, which is pre-trained on medical corpora. The MTC task is performed following a pre-training task MTS instead of adding additional classifiers as previous methods did. Specifically, we wrap the raw text into a new input sequence with a template and calculate the probability distribution of candidate words corresponding to the [UNK]. The category of a medical text can be inferred by the predicted word without using extra parameters or classifiers. Finally, the experimental results show that our method outperforms the benchmark and previous approaches on both KUAKE-QIC and CHIP-CTC datasets.

However, there are still potential challenges and opportunities for future research. Firstly, our method still relies on the quality and quantity of training data. Insufficient or noisy training data may negatively affect the performance. Secondly, our method may not be robust to noise and errors in the input. In the real world, the input text may contain errors or noise, which may affect the performance. Lastly, the generalizability of our model to other languages and domains is an open research question. In our experiments, we focused on the Chinese language and the medical domain. However, the effectiveness of our model in other languages and domains remains to be explored. For the future work, we can further explore the importance of domain-specific pre-training and prompt engineering for other NLP tasks based on the proposed method, or explore the generalizability of our method to other languages and domains.

Acknowledgements: This research was supported by the Medical Big Data Supercomputing Center System of Anhui Medical University.

Contributorship: YW (Yu Wang) was the lead in conceptualization, methodology, and initial manuscript draft writing. YW (Yuan Wang) conducted the literature review. ZP and FZ prepared the dataset. ZY contributed to code implementation. FY was the lead in funding acquisition and supervision. The manuscript was reviewed and edited by all authors, and the final version was approved.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study not involve primary data collection, and formal ethics approval will therefore not be required.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Natural Science Foundation of Anhui Province of China (Nos. 2108085MH303 and 2108085QF274) and the Initiation Fund of Anhui Medical University (No. 1401039201).

Guarantor: FY.

Informed consent: Patient consent is not required because the research does not involve primary data collection and we are using publicly available medical text datasets that do not contain private information.

ORCID iDs: Yu Wang https://orcid.org/0000-0003-3096-2481
Feifan Zhang https://orcid.org/0000-0003-4551-0365

Notes

1. https://github.com/paddlepaddle/paddle
2. This study utilized publicly available datasets without any private information and no data collection was conducted.

References

1. Richter-Pechanski P, Geis NA, Kiriakou C et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. *Digital Health* 2021; 7: 20552076211057662.
2. Altman R. Artificial intelligence (AI) systems for interpreting complex medical datasets. *Clin Pharmacol Ther* 2017; 101: 585–586.
3. Névéol A, Dalianis H, Velupillai S et al. Clinical natural language processing in languages other than English: Opportunities and challenges. *J Biomed Semantics* 2018; 9: 1–13.
4. Saad E, Sadiq S, Jamil R et al. Predicting death risk analysis in fully vaccinated people using novel extreme regression-voting classifier. *Digital Health* 2022; 8: 20552076221109530.
5. Mujtaba G, Shuib L, Idris N et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl* 2019; 116: 494–520.
6. Sarker A and Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform* 2015; 53: 196–207.
7. Yi K and Beheshti J. A hidden Markov model-based text classification of medical documents. *J Inform Sci* 2009; 35: 67–81.
8. Yahia HS, Abdulazeez AM et al. Medical text classification based on convolutional neural network: a review. *Int J Sci Bus* 2021; 5: 27–41.
9. Lavanya P and Sasikala E. Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: a comprehensive survey. In: *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. IEEE, pp.603–609.
10. JDM W C Kenton and Toutanova LK. Bert: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*. pp.4171–4186.
11. Sun Y, Wang S, Li Y et al. Ernie 2.0: a continual pre-training framework for language understanding. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.8968–8975.
12. Clark K, Luong MT, Le QV et al. ELECTRA: pre-training text encoders as discriminators rather than generators. In: *Proceedings of ICLR*, pp.1–18.
13. Petroni F, Rocktäschel T, Riedel S et al. Language models as knowledge bases? In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.2463–2473.
14. Davison J, Feldman J and Rush AM. Commonsense knowledge mining from pretrained models. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp.1173–1178.
15. Qasim R, Bangyal WH, Alqarni MA et al. A fine-tuned bert-based transfer learning approach for text classification. *J Healthc Eng* 2022; 2022.
16. Guo Y, Ge Y, Yang YC et al. Comparison of pretraining models and strategies for health-related social media text classification. *Healthcare* 2022; 10: 1478.
17. Hu S, Ding N, Wang H et al. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.2225–2240.
18. Schick T and Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp.255–269.
19. Wang Q, Dai S, Xu B et al. Building chinese biomedical language models via multi-level text discrimination. *arXiv preprint arXiv:211007244* 2021.
20. Goodfellow I, Pouget-Abadie J, Mirza M et al. Generative adversarial networks. *Commun ACM* 2020; 63: 139–144.
21. Minaee S, Kalchbrenner N, Cambria E et al. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)* 2021; 54: 1–40.
22. Zhang N, Chen M, Bi Z et al. Cblue: A chinese biomedical language understanding evaluation benchmark. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.7888–7915.
23. Koopman B, Zuccon G, Nguyen A et al. Automatic ICD-10 classification of cancers from free-text death certificates. *Int J Med Inform* 2015; 84: 956–965.
24. Kocbek S, Cavedon L, Martinez D et al. Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. *J Biomed Inform* 2016; 64: 158–167.
25. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM* 2017; 60: 84–90.
26. Rios A and Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In: *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pp.258–267.
27. Hughes M, Li I, Kotoulas S et al. Medical text classification using convolutional neural networks. In: *Informatics for*

Health: Connected Citizen-Led Wellness and Population Health, IOS Press, 2017, pp.246–250.

28. Nii M, Tsuchida Y, Kato Y et al. Analysis of classification results for the nursing-care text evaluation using convolutional neural networks. In: *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, IEEE, pp.1–6.
29. Yao L, Mao C and Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 2019; 19: 31–39.
30. Oleynik M, Kugic A, Kasáč Z et al. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *J Am Med Inform Assoc* 2019; 26: 1247–1254.
31. Venkataraman GR, Pineda AL, Bear Don't Walk IVOJ et al. Fastag: automatic text classification of unstructured medical narratives. *PLoS ONE* 2020; 15: e0234647.
32. Liang S, Chen X, Ma J et al. An improved double channel long short-term memory model for medical text classification. *J Healthc Eng* 2021; 2021.
33. Bangyal WH, Qasim R, Ahmad Z et al. Detection of fake news text classification on COVID-19 using deep learning approaches. *Comput Math Method Med* 2021; 2021.
34. Zhou X, Li Y and Liang W. CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Trans Comput Biol Bioinform* 2020; 18: 912–921.
35. Li X, Cui M, Li J et al. A hybrid medical text classification framework: integrating attentive rule construction and neural network. *Neurocomputing* 2021; 443: 345–355.
36. Ibrahim MA, Khan MUG, Mehmood F et al. GHS-net a generic hybridized shallow neural network for multi-label biomedical text classification. *J Biomed Inform* 2021; 116: 103699.
37. Shin B, Chokshi FH, Lee T et al. Classification of radiology reports using neural attention models. In: *2017 international joint conference on neural networks (IJCNN)*, IEEE, pp.4363–4370.
38. Deng J, Dong W, Socher R et al. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp.248–255.
39. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 6000–6010.
40. Lu H, Ehwerhemuepha L and Rakovski C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med Res Methodol* 2022; 22: 1–12.
41. Peng Y, Yan S and Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, pp.58–65.
42. Gao S, Alawad M, Young MT et al. Limitations of transformers on clinical text classification. *IEEE J Biomed Health Inform* 2021; 25: 3596–3607.
43. Jiang G, Liu S, Zhao Y et al. Fake news detection via knowledgeable prompt learning. *Inf Process Manag* 2022; 59: 103029.
44. Li H, Mo T, Fan H et al. Kipt: knowledge-injected prompt tuning for event detection. In: *Proceedings of the 29th International Conference on Computational Linguistics*. pp.1943–1952.
45. Wei L, Li Y, Zhu Y et al. Prompt tuning for multi-label text classification: How to link exercises to knowledge concepts? *Appl Sci* 2022; 12: 10363.
46. Cui Y, Che W, Liu T et al. Pre-training with whole word masking for Chinese BERT. *IEEE Transactions on Audio, Speech and Language Processing*, 2021.
47. Cui Y, Che W, Liu T et al. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.657–668.
48. Zhuang L, Wayne L, Ya S et al. A robustly optimized BERT pre-training approach with post-training. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pp.1218–1227.

Appendix

The appendix provides the details of the random experiments mentioned in the "Results of fine-tuning and prompt-tuning" section. Table 7 displays the results of fine-tuning and prompt-tuning using BERT and ERNIE-Health on the KUAKE-QIC dataset, including the accuracy values, *p*-values of Shapiro–Wilk tests and *T*-tests. Additionally, Table 8 presents the experimental results on the CHIP-CTC dataset.

Based on the experimental findings, it can be observed that the results obtained from eight random experiments satisfy the normality assumption, as evidenced by the *p*-values of Shapiro–Wilk tests exceeding the significance level of 0.05. Consequently, paired *T*-tests are conducted to examine whether the prompt-tuning yields significantly improvement compared to fine-tuning. From the findings presented in Table 7, it is evident that for the KUAKE-QIC dataset, irrespective of whether BERT-base or ERNIE-Health is employed, the *p*-values derived from the *T*-test are below the significance level of 0.05. Thus, it can be concluded that prompt-tuning demonstrates significantly improvement relative to fine-tuning on this particular dataset. Conversely, as illustrated in Table 8, for the CHIP-CTC dataset, when ERNIE-Health is utilized, the *p*-value derived from the *T*-test falls below the significance level of 0.05. However, when BERT-base is employed, the *p*-value marginally surpasses 0.05. Consequently, given that the training sample size of the CHIP-CTC dataset is three times greater than that of the KUAKE-QIC dataset, it can be suggested that prompt-tuning may exhibit more favorable performance on datasets with smaller-scale training samples.