ISSN 1330-3651 (Print), ISSN1848-6339 (Online)

https://doi.org/10.17559/TV-20240721001871
Original scientific paper

Regional Disparities and Evolutionary Trends in Physician Resources in China: A Deep Learning and Text Mining Approach

Siyu WANG, Mingyang LI*, Limin WANG*, Xuming HAN, Jiawei LI

Abstract: Physician resources are fundamental assets of the healthcare system. The unequal distribution of these resources inevitably leads to patients seeking medical services elsewhere, increasing healthcare costs, and hindering the realization of health equity. However, the phenomenon of uneven healthcare resource distribution is widespread, deviating from the goal of balanced regional development. Moreover, few studies have explored regional disparities and evolutionary trends in physician resources from the perspective of doctors' professional skills. To address this gap, this paper proposes a method for assessing regional differences in physician resources based on text big data, from the perspective of doctors' professional skills. The method uses a deep learning-based named entity recognition approach to extract two types of relevant entities diseases and treatment methods and assigns weights to these entities using their inverse document frequency index. The effectiveness of this method is validated through experiments, and based on large-scale online medical community text data. The proposed method is used to explore the provincial disparities and their evolutionary trends over time in physician resources in China. The results indicate that physician resources exhibit significant spatial and temporal heterogeneity. Professional skills should be taken into consideration in order to achieve a balanced, rational, and efficient distribution of physician resources.

Keywords: deep learning; evolutionary trends; physician resources; regional disparities; text mining

1 INTRODUCTION

As a vital component of public services, the equitable distribution of medical resources significantly influences residents' access to medical care and is intrinsically linked to the healthy and sustainable progression of society [1]. China's ongoing healthcare reforms have consistently improved the nation's medical service system, markedly enhancing the quality of healthcare services [2]. Nevertheless, regional economic disparities, healthcare financing imbalances, resident income inequality, and segmentation in medical insurance contribute to the inequitable allocation of medical resources among different geographic areas [3].

A disproportionate concentration of high-quality medical resources-encompassing cutting-edge technologies, top-tier medical professionals, and advanced equipment-is observed in urban centers, leaving rural areas at a disadvantage [4]. This distribution inequality has significant ramifications. It leads to increased inter-regional patient migration, elevated healthcare costs, and a departure from the principle of health equity and the aim of harmonious regional development [5]. In response, the National Development and Reform Commission of China has launched initiatives to mitigate these disparities. These include the selection of leading medical facilities in resource-rich areas to establish branches or sub-centers and promoting multiple practice sites for doctors. Such measures aim to bridge the treatment efficacy gap between provinces with healthcare resource shortages and their more affluent counterparts, thereby reducing inter-provincial patient mobility [6]. This backdrop raises pertinent questions: How have physician resources in China evolved? What regional variances exist? Addressing these questions is not only academically significant but also pivotal for healthcare system reform and the equitable distribution of high-quality medical resources.

The continuous development of deep learning technology has promoted the application of text mining technology in the medical field. By analyzing large volumes of unstructured data from medical literature, electronic medical records, patient reports, clinical records, and other related documents, this technology helps physicians, researchers, and medical institutions obtain valuable information to improve diagnostic accuracy, optimize treatments, and advance medical research [7, 8]. At the same time, the accumulation and diversification of healthcare data have enriched unstructured text data such as electronic medical records, patient feedback, medical reports, hospital logs, and doctor-patient interactions from online healthcare communities. This enables data mining based on large-scale text to help managers and decision makers make more scientific and reasonable resource allocation decisions [9]. However, a review of existing studies shows that most research has focused on the regional disparities in physician resources and the optimization of physician resource allocation under the situation of medical reform. Most of these studies take the macro data of physician resources as experimental data, and use different indicators and models to measure disparities in physician resources across different periods and regions. However, few studies have explored the differences of doctors' professional skills in different regions and the evolutionary trends over time from the perspective of doctors' professional skills. The named entity recognition method based on deep learning in text mining can accurately and efficiently extract the diseases and treatment methods that physicians specialize in from the text, so as to conduct a more in-depth analysis of physicians' professional skills [10-12].

With the increasing convergence of the Internet and healthcare services, online medical care has become an alternative to conventional care [13, 14]. Online medical communities leverage internet technologies to connect patients, doctors, and hospitals. This facilitates access to services such as consultations, evaluations, registrations, and health advice [15, 16]. "Haodaifu Online," one of China's largest doctor-patient platforms, includes over 10000 hospitals and nearly 900000 doctors, reflecting the offline distribution of physicians regionally [17]. Doctors on this platform fall into two categories: those with personal pages providing consultations and those listed by the platform without personal pages. Nevertheless, the data on "Haodaifu Online" is a reliable indicator of a doctor's expertise, which, in turn, signifies the medical standard of the region [18].

Accordingly, this study utilizes "Haodaifu Online" data, extracted via Python in the Jupyter programming environment, to analyze specialty text data from registered doctors as of October 2022. A BERT-BiLSTM-CRF model is constructed for named entity recognition (NER) to discern entities related to diseases and treatments from the specialty texts. After preprocessing and annotating the text data for training and testing, the model's performance is compared with classical methods, underscoring its effectiveness. Moreover, the study proposes methods to evaluate doctor expertise regionally and temporally, employing spatial visualization to examine regional variances in expertise levels. It also investigates the evolution of doctor expertise over time within online medical communities. This research contributes to understanding the spatial and temporal heterogeneity of medical resource distribution in China.

The remaining sections are arranged as follows: Section 2 examines pertinent scholarly works. The data and technique are explained in Section 3. The findings and discussion are presented in Section 4. The implications for policy are discussed in Section 5.

2 RELATED WORK

2.1 Named Entity Recognition for the Medical Field

Named entity recognition stands as a pivotal area within natural language processing (NLP), focusing on the identification and categorization of textual entities [19, 20]. In the medical domain, the text is often unstructured and abundant with terminology such as diseases, symptoms, and drugs. The objective is to mine this data to uncover underlying patterns and knowledge, necessitating precise NER for entities and their interrelationships [21-23]. Rule-based and dictionary-based strategies have given way to statistical techniques for machine learning as well as deep learning in the development of NER methods, some researchers have combined these approaches for increased effectiveness [24-27].

Chinese language structure differs significantly from Indo-European languages, presenting challenges in entity boundary recognition [28]. The nuances of Chinese word formation, character polysemy, and textual complexity pose obstacles for effective Chinese NER [29]. However, the field, particularly in the medical context, is advancing rapidly [30]. Lei et al. investigated the effects of different types of feature including bag-of-characters, word segmentation, part-of-speech, and section information, and different machine learning algorithms including conditional random field (CRF), support vector machine (SVM), maximum entropy, and structural SVM on the Chinese clinical NER task [31]. Zhao et al. introduced an adversarial training-based lattice Long Short-Term Memory (LSTM) with a CRF layer for Chinese clinical NER [32]. Furthermore, Qin et al. developed a BERT-BiGRU-CRFNER method for Chinese electronic medical records, focusing on cerebrovascular diseases to extract critical entity categories such as disease, symptom, body part, medical examination, and treatment [33].

In general, there are four NER methods in the medical field: traditional rule-based methods, statistical learning methods, deep learning methods, multi-task learning methods, and transfer learning methods. Among these, the latter two methods have become the current mainstream. Some researchers have also suggested integrating various techniques into a single method. Different NER methods are suited for distinct application scenarios depending on their specific features and advantages. The medical field involves a large number of technical terms, abbreviations, synonyms, and emerging terms, which place higher demands on the NER model. Currently, no single NER model can be universally applied to all scenarios.

2.2 Spatial and Temporal Characteristics Analysis of Physician Resources

The equitable and logical distribution of physician resources is crucial for public access to healthcare services [34]. Despite this, the global challenge of physician shortages and their maldistribution persists [35]. Researchers have explored these disparities at various spatial scales. For identifying regions with physician shortages, Xiong et al. offered a geographic information system (GIS)-based proximate area method and gravity method [36]. Erdenee et al. applied the Gini coefficient to reveal that, in Mongolia, the per capita distribution of doctors was fair, but significant disparities emerged when considering distribution per unit area [37]. Beyond regional assessments, temporal analyses have been conducted to understand the evolution of physician distribution. Paramita et al. used the Gini index to track physician distribution trends across Indonesia's 34 provinces from 2000 to 2014 [38]. As one of the indicators to assess the quality of health resources, Wang et al. computed the Gini coefficient and Concentration index using the amount of physicians with varying levels of education [39]. In conclusion, they found that the proportion of physicians overall and physicians holding bachelor's degrees or higher in primary health care institutions in Liaoning Province, China, decreased after 2011. This finding suggests that the efforts to enhance the educational attainment of physicians in primary health care institutions have been fruitful. Pal et al. employed descriptive statistics method to examine the geographic and temporal distribution of physicians in the European Union from 2006 to 2018 [40]. Yan et al. demonstrated the growing divergence in physician and healthcare bed distribution across China's prefecture-level cities from 2000 to 2018 by employing Gini coefficients as well as Moran's Index in bivariate form [41].

Most spatial-temporal analyses have utilized statistical methods like descriptive statistics, concentration indices, and Gini coefficients to calculate physician resource allocation by population or geographic area [42, 43]. These approaches generally consider the number of physicians without addressing the complexity of the medical profession. However, the professional skills of physicians are critical for enhancing medical resource efficiency, providing quality healthcare services, and minimizing medical errors and disputes. Hence, assessing physician resources should encompass the professional skills of doctors beyond mere headcounts [37].

3 DATA AND METHODOLOGY

3.1 Methodological Flow

The study's framework for examining regional disparities and evolutionary trends in doctor resources is outlined in Fig. 1. NER, a technique derived from text mining, is used to extract entities related to diseases and treatments from the professional skill texts of doctors in online medical communities. The research evaluates disparities in these categories of entities across various regions. It also assesses how these disparities change over time, correlating with the doctors' duration of activity on the platform. The methodology involves three primary phases: data collection, data processing, and data analysis.

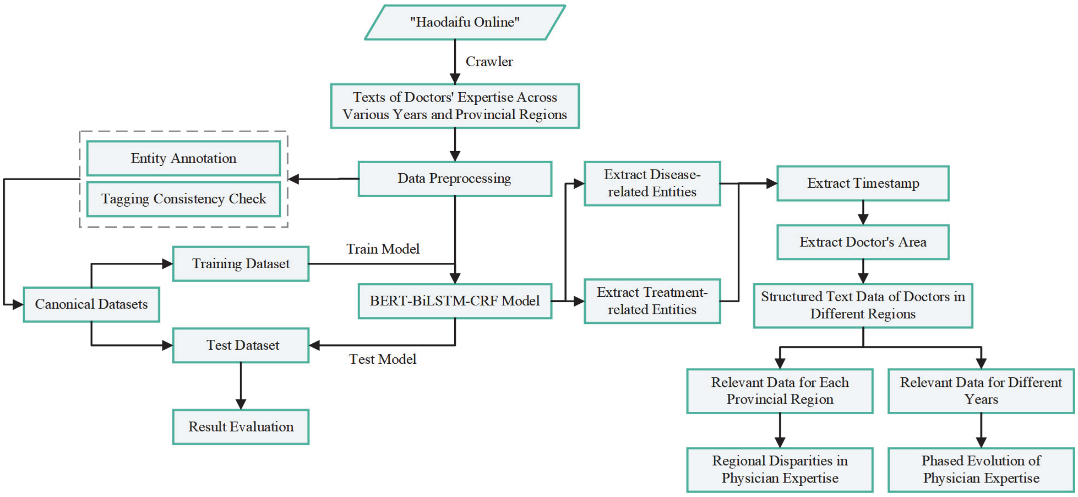

"Haodaifu Online"
Crawler
Texts of Doctors' Expertise Across Various Years and Provincial Regions
Entity Annotation
Tagging Consistency Check
Data Preprocessing
Train Model
Training Dataset
Canonical Datasets
BERT-BiLSTM-CRF Model
Test Dataset
Test Model
Result Evaluation
Extract Disease-related Entities
Extract Treatment-related Entities
Extract Timestamp
Extract Doctor's Area
Structured Text Data of Doctors in Different Regions
Relevant Data for Each Provincial Region
Relevant Data for Different Years
Regional Disparities in Physician Expertise
Phased Evolution of Physician Expertise

Figure 1 The methodological flow

3.2 Data

Physician expertise is defined by the capacity of a doctor to independently make diagnoses and provide treatment of specific diseases, relying on an extensive medical knowledge base and practical experience. Given the distinct object-oriented and practical realms of doctor specialties, significant variances in expertise are expected among doctors from various departments. For a focused and practical analysis, this study concentrates on the cardiovascular medicine department, sourcing public text data on cardiovascular physicians' expertise from the "Haodaifu Online" platform to evaluate provincial differences and the evolutionary trends of physician resources. Cardiovascular diseases are a major health threat to residents in China, with characteristics like high prevalence, abundant data, and easy accessibility. Studying physician resources related to cardiovascular diseases can reflect common issues in the distribution of medical resources across China [44]. The methodology implemented here can be extended to similar research in other medical departments.

Table 1 Number of physicians in provinces, autonomous regions and municipalities

Provinces, Autonomous Regions, Municipalities	Number of Physicians	Provinces, Autonomous Regions, Municipalities	Number of Physicians	Provinces, Autonomous Regions, Municipalities	Number of Physicians
Beijing	1583	Shandong	1010	Qinghai	36
Hebei	442	Henan	695	Neimeng	100
Shanxi	248	Hubei	683	Guangxi	117
Liaoning	404	Hunan	373	Xizang	13
Jilin	179	Guangdong	909	Ningxia	121
Heilongjiang	276	Hainan	53	Xinjiang	292
Jiangsu	765	Sichuan	439	Shanghai	840
Zhejiang	588	Guizhou	83	Tianjin	303
Anhui	276	Yunnan	329	Chongqing	194
Fujian	325	Shanxi	364		
Jiangxi	242	Gansu	96		

Data on cardiovascular physicians from various Chinese regions was collected via web crawling from "Haodaifu Online," including doctors' specialty texts, affiliated hospitals, hospital locations, and the dates of the doctors' account creations on the platform. The initial dataset comprised 12621 entries of doctors' text data. After refining this data to exclude entries with missing expertise information or unclear provincial locations, a final dataset of 10431 valid text entries remained, spanning hospitals in cities of varying sizes nationwide. The timeframe of platform account creations ranged from 2011 to 2021. The distribution of physicians among different provinces, autonomous regions, while municipalities is presented in Tab. 1, with the exception of Macao, Taiwan Province, as well as Hong Kong. Additionally, the discrepancies in the amount of physicians across these municipalities, autonomous regions, and provinces are illustrated graphically in Fig. 2

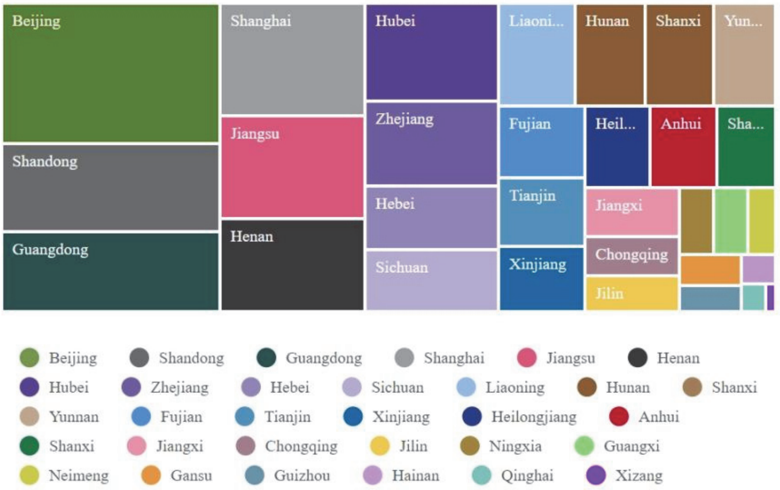


Figure 2 The tree-like distribution of the number of doctors' information texts in provinces, autonomous regions and municipalities

3.3 Methods

The current research on NER in medical texts predominantly addresses electronic medical records, medical literature, and medical books, with limited focus on texts from online medical communities [45, 46]. It is only recently that the latter has garnered research attention [47]. NER methods in the medical field confront challenges, including the poor quality of annotated data, reliance on a single text feature, overlooking text dependencies, and suboptimal recognition outcomes [48, 49]. To overcome these issues, this study introduces a deep learning hybrid model that integrates the pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model with Bidirectional Long Short-Term Memory (BiLSTM) as well as CRF for NER in online medical community texts. In this BERT-BiLSTM-CRF model, BERT provides high-quality feature representations and strong semantic information extraction capabilities; BiLSTM processes contextual information from text sequences to enhance feature extraction; and CRF utilizes the context of adjacent labels to ensure the optimal global label for each entity. Fig. 3 illustrates the BERT-BiLSTM-CRF model's architecture.

3.3.1 BERT Approach

The BERT model is a significant advancement in language representation, introduced by Devlin et al. from Google in October 2018. It has set new benchmarks in 11 NLP tasks due to its impressive capabilities. Combining the strengths of the ELMo and OpenAIGPT models, BERT utilizes the Transformer structure as its core to perform truly bidirectional training, integrating both preceding and following context in its predictions [50]. The architecture of BERT is detailed in Fig. 4. BERT employs a two-stage training process. The first stage involves feature extraction using the Transformer encoder and pre-training on the Chinese Wikipedia corpus through two unsupervised prediction tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM operates at the word level, masking a percentage of words randomly for the model to predict, thus creating a deep bidirectional representation. NSP, at the sentence level, is a binary classification task that discerns whether two sentences are consecutive, leveraging the relational features between sentences. Both tasks utilize the cross-entropy loss function, with the cumulative loss that is equal to the sum of the individual losses from these tasks. After extensive pre-training using a sizable corpus, such model parameters obtained are then applied to downstream tasks. These pre-trained parameters are fine-tuned with supervision using the task's training corpus, thus completing the second training stage for the BERT model.

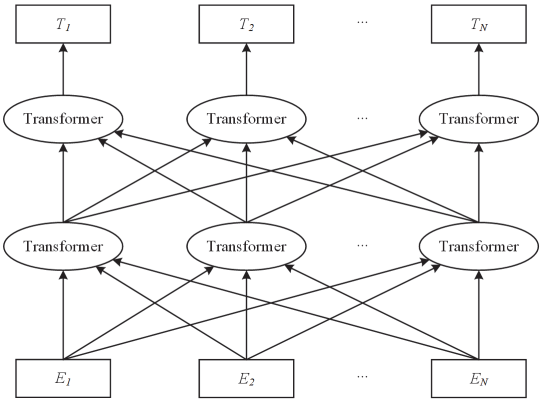


Figure 4 Schematic diagram of the BERT model

Position Embedding, Token Embedding, and Segment Embedding are the three components of the input that generate BERT. These represent the word, sentence, and positional information, respectively. Token Embeddings are word representations derived from unsupervised training, where semantically similar words have similar vectors. Segment Embeddings mark whether the input word belongs to sentence A or B, differentiating sentences or paragraphs within the text. Position Embeddings encode each word's position within the sentence, incorporating positional information into the text. The BERT model enhances the generalizability of word vectors, enabling deep learning of features at character, word, and sentence levels and even between sentences. It dynamically adjusts word vectors based on context, effectively addressing the challenge of polysemy that static word vectors cannot resolve and outperforming previous methods. Therefore, this study employs BERT for text vectorization conversion in the pre-training model.

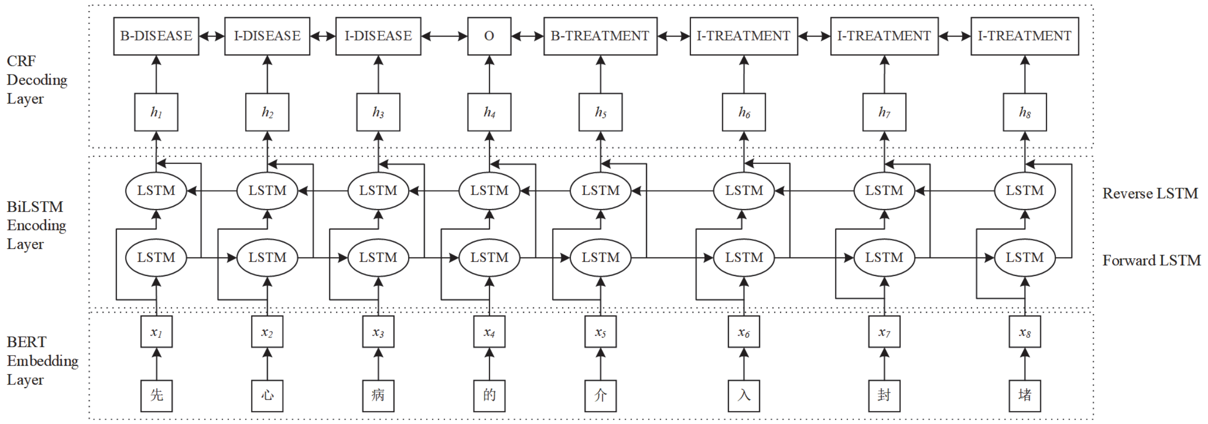


Figure 3 BERT-BiLSTM-CRF framework and its application in Chinese named entity recognition

3.3.2 BiLSTM Approach

The emergence of deep learning has led many researchers to employ deep neural networks for NER. The key advantage of this methodology is its independence from manual data preprocessing, with the network itself capable of feature extraction through training [51]. Recurrent neural networks (RNNs), especially those based on a deep structure, are the most commonly used deep learning algorithm for NER. Traditional neural networks fail to process sequential data with temporal characteristics, like natural language. RNNs address this limitation by incorporating loops that feed the network's output back into itself, allowing the network to make use of sequential input features. RNNs can theoretically use historical information for predictions, which is particularly beneficial for text processing tasks that involve classified targets. However, they face challenges when the sequence lengthens, as gradients can diminish exponentially, causing a loss of the ability to learn from longer-distance data. To counter this, LSTM network was first introduced by Hochreiter and Schmidhuber in 1997, an RNN variant that resolves the issues of vanishing and exploding gradients, thus retaining more historical information [52]. While unidirectional RNNs can only capture historical sequence information, labeling tasks often require context in both directions. To incorporate this contextual information, Graves and Schmidhuber developed the BiLSTM network, which extends the unidirectional LSTM into a bidirectional structure [53]. BiLSTM networks consist of forward and backward LSTM networks that allow for bidirectional information flow.

LSTM is a variant of RNN, and all RNNs have a chain structure of repetitive neural network modules. In the basic RNN, repetitive modules are just very simple structures, such as a tanh layer. However, in the LSTM model, each repetitive neuron contains three gate structures (forgetting gate, input gate and output gate) to protect and control the information state. Their roles are respectively to selectively forget part of the historical information, add part of the current input information, and finally integrate the current state and produce output. The specific structure of LSTM is illustrated in Fig. 5, where the variable c represents the memory information stored by the cell and propagates throughout the entire model. In contrast to RNN, besides the output h that flows over time, in the LSTM model, cell state c also flows with time, which represents long-term memory. The memory unit is responsible for the storage of historical information, recording and updating historical information through a state parameter; The gate structure determines the trade-off of information through the sigmoid function, and thus acts on the memory unit.

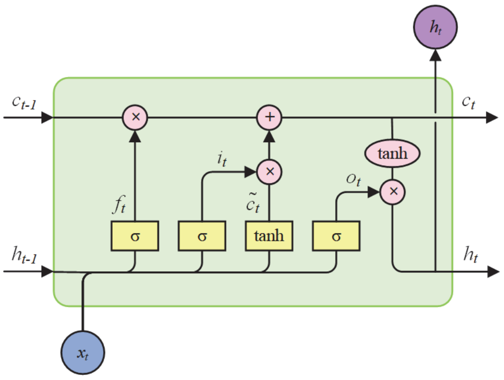


Figure 5 Schematic diagram of the LSTM unit structure

In the LSTM model, the initial step involves computing the forgetting gate, which filters information to be discarded. This computation uses the current timestep data, x_t, and the output from the last timestep, h_{t-1}. A sigmoid activation function scales the previous cell state, c_{t-1}, within a range of 0 to 1. A value of 0 indicates the complete omission of prior cell state information, while a value of 1 indicates full preservation. The forgetting gate's formula is depicted in Eq. (1).

$$f_t = \sigma\left(W_f \times \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

The subsequent step in the LSTM process entails the computation of the input gate, which identifies information to be stored. The inputs remain x_t and h_{t-1}. The input gate's sigmoid function ascertains the values to be updated, and concurrently, a candidate vector is produced via the input gate's tanh layer, serving as a filter for newly acquired information. This is algebraically articulated in Eq. (2) and Eq. (3). Thereafter, the cell state, $\tilde{c}_t$, is updated according to the methodology outlined in Eq. (4).

$$i_t = \sigma\left(W_i \times \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\tilde{c}_t = \tan h\left(W_c \times \left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{4}$$

In the third step, the model calculates the output gate and uses it alongside the refreshed cell state to calculate the current LSTM neuron hidden layer output. The output gate selects which cell state features to express based on h_{t-1} and x_t, as demonstrated in Eq. (5). The approach to calculating the current hidden layer output of LSTM neurons is encapsulated in Eq. (6).

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t \times \tan h\left(c_t\right) \tag{6}$$

Within the equation, W and b represent the weights and biases associated with the gates, and σ symbolizes the sigmoid function, which was selected for the nonlinear transformations required in this study.

Research within the realm of NLP often necessitates the balanced consideration of preceding and subsequent information during the training of models. This is particularly pertinent in sequence labelling tasks such as NER. While the LSTM model is adept at processing long-term relational data and addressing issues related to long-distance dependencies, its capability to fully utilize contextual information in a simultaneous manner is limited. Building upon the principles of LSTM, the BiLSTM model is employed to overcome these limitations. BiLSTM harnesses the semantic features of the text from both directions: the forward LSTM layer extracts features from the beginning to the end of the text (as outlined in Eq. (7)), while the reverse LSTM layer does so from the end to the beginning (refer to Eq. (8)). By amalgamating these two sets of features, BiLSTM provides a comprehensive understanding of the contextual text features (see Eq. (9)). Accordingly, this study applies the BiLSTM model to extract pivotal features for NER in Chinese physicians' online community profiles.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}\left(x_t\right) \tag{7}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}\left(x_t\right) \tag{8}$$

$$h_t =< \overrightarrow{h}_t, \overleftarrow{h}_t > \tag{9}$$

In the aforementioned equations, $\overrightarrow{h_t}$ symbolizes the hidden state derived from the forward LSTM layer, whereas $\overleftarrow{h_t}$ represents the hidden state from the reverse LSTM layer. $\overrightarrow{LSTM}\left(x_t\right)$ indicates the feature representation for sequential text analysis from front to back, and $\overleftarrow{LSTM}\left(x_t\right)$ denotes the feature representation for analysis from back to front. Finally, ht epitomizes the ultimate hidden layer state, which is the amalgamation of vectors from both the forward and reverse analyses.

3.3.3 CRF Approach

The CRF model is a discriminative probabilistic undirected graph model suited for sequence labelling tasks. It computes the conditional probability of an entire labelled sequence, given an observation sequence. Uniquely, CRF amalgamates features of both the maximum entropy model and the hidden Markov model. It surpasses the hidden Markov model by not adhering to the independent observation hypothesis and the homogeneity condition of Markov processes. Furthermore, CRF overcomes the annotation bias issue found in the maximum entropy Markov model by globally normalizing input features, resulting in a globally optimal solution. This study integrates CRF with the NER model to assimilate contextual information and semantic links, thereby boosting the precision of NER. The most prevalent CRF variant employed here is the linear chain CRF, illustrated in Fig. 6. This figure represents variables as nodes and their dependencies as edges. Eq. (10) and Eq. (11) detail the conditional probability of the linear chain CRF.

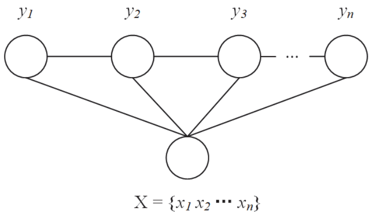


Figure 6 Graph structure of linear chain conditional random field

$$P\left(\mathrm{y} \mid x\right) = \frac{1}{Z\left(x\right)} \exp\left(\begin{array}{l} \sum_{i=1}^{n-1}\sum_{k} \lambda_k f_k\left(y_{i+1}, y_i, x, i\right) + \\ + \sum_{i=1}^{n}\sum_{l} \eta_l g_{\,l}\left(y_i, x, i\right) \end{array}\right) \tag{10}$$

$$Z\left(x\right) = \sum_{y} \exp\left(\begin{array}{l} \sum_{i=1}^{n-1}\sum_{k} \lambda_k f_k\left(y_{i+1}, y_i, x, i\right) + \\ + \sum_{i=1}^{n}\sum_{l} \eta_l g_{\,l}\left(y_i, x, i\right) \end{array}\right) \tag{11}$$

where $f_k\left(y_{i+1}, y_i, x, i\right)$ is a transfer feature function specified at two label positions that are adjacent in the observation sequence. The correlation between adjacent label variables and their impact on the observation sequence is delineated by this function. $g_l\left(y_i, x, i\right)$ is a state feature function at the labelled position i of the observation sequence, elucidating the impact of the observation sequence on the label variable. Both the transfer and state feature functions are real-valued, typically represented by 0 and 1, indicating the empirical likelihood of corresponding labels y_i and y_{i+1} given the i-th observation x_i. λ_k and η_l are the parameters, k and l represent the number of transfer and state functions, respectively, and $Z(x)$ represents the normalization factor.

In viewing NER as a sequence labelling problem, a sequential order between adjacent labels is observed. For instance, the 'I-Disease' label (indicating the middle of a disease entity) follows the 'B-Disease' label (signifying the beginning of the disease entity). BiLSTM maintains separate hidden states for forward and reverse recurrent neural networks. The forward LSTM transmits its hidden state solely to the next forward LSTM, and likewise for the reverse LSTM. No interaction occurs between the neural networks in both directions, and their outputs converge only at the output node to synthesize the final output. Sole reliance on the BiLSTM model for entity recognition may lead to inadequate representation of the constraint relationships between annotations. Therefore, this study appends a CRF layer following the BiLSTM layer, transforming the hidden status sequence $h=\{h_1, h_2, ..., h_t\}$ into its optimal labelling sequence $y=\{y_1, y_2, ..., y_t\}$.

The integration of BiLSTM and CRF is fundamental to this research. This method involves utilizing the output of the BiLSTM as the input for the CRF. First, the BiLSTM extracts sequence features, and then the CRF is employed to train sentence-level label information. More concretely, a linear layer is appended to the BiLSTM's output layer, and the hidden layer output result (feature vector) generated by BiLSTM is converted into the score corresponding to each label by linear transformation, assuming *P* as the score matrix for the input sentence $X=(x_1, x_2, ..., x_n)$ through the BiLSTM layer. The dimensions of *P* are $n \times k$, where *n* is the number of words in the sentence and *k* is the number of tags, and P_{ij} signifies the probability (score) that the *i*-th word in the sentence is labelled as the *j*-th tag. For a given input sentence *X*, a predictive tag sequence $y=(y_1, y_2, ..., y_n)$ is derived. The score function for this sequence is outlined in Eq. (12).

$$s(X, y)=\sum_{i=0}^{n} A_{y_i, y_{i+1}}+\sum_{i=1}^{n} P_{i, y_i} \tag{12}$$

In Eq. (12), A symbolizes the transfer score matrix, $A_{y_i, y_{i+1}}$ indicates the transferred score from label y_i to y_{i+1}. The labels y_0 and y_{n+1} represent the start and end labels of the sentence, respectively, and the dimensions of A are $(k+1)\times(k+1)$. The probability $P(y \mid X)$ of the predicted sequence *y*, given the sentence *X*, is calculated as per Eq. (13), where Y_X denotes all potential labelled sequences for a given sentence *X*. During the decoding phase, the Viterbi algorithm is applied to identify the highest-scoring sequence y^* from all possible labelled sequences, as described in Eq. (14). This process ensures the selection of the globally optimal labelled sequence.

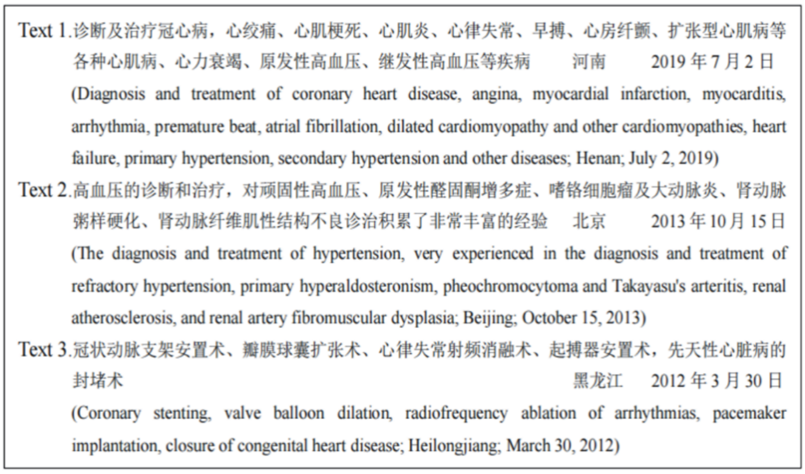
$$P(y \mid X)=\frac{\exp(s(X, y))}{\sum_{\tilde{y} \in Y_X} \exp(s(X, \tilde{y}))} \tag{13}$$

$$y^*=\underset{\tilde{y} \in Y_X}{\operatorname{argmax}}\, s(X, \tilde{y}) \tag{14}$$

4 RESULTS AND DISCUSSION

4.1 Data Preprocessing and Entity Annotation

The initial phase of this research involved data collection, sourced through web crawling, and the subsequent elimination of entries with missing values. This process yielded a total of 10431 text records pertaining to cardiovascular physicians. Fig. 7 exemplifies the format of this text data. Each record encompasses details such as the physician's specialty, the provincial region of their hospital affiliation, and the date their profile was created on the platform. The collected data are primarily semi-structured, aligning closely with the official public information from hospitals. They exhibit characteristics like uniform formatting and minimal noise data. As a result, there was no necessity for noise reduction or stop word removal during the preprocessing phase. This study employs a character-level NER approach, thereby negating the need for word segmentation in the text. The first 1000 text entries were selected for both training and testing purposes, maintaining a 3:1 ratio between the two. For the entity annotation process, each word in the training and test data was isolated into a separate row.



Text 1.诊断及治疗冠心病，心绞痛、心肌梗死、心肌炎、心律失常、早搏、心房纤颤、扩张型心肌病等各种心肌病、心力衰竭、原发性高血压、继发性高血压等疾病　河南　2019年7月2日

(Diagnosis and treatment of coronary heart disease, angina, myocardial infarction, myocarditis, arrhythmia, premature beat, atrial fibrillation, dilated cardiomyopathy and other cardiomyopathies, heart failure, primary hypertension, secondary hypertension and other diseases; Henan; July 2, 2019)

Text 2.高血压的诊断和治疗，对顽固性高血压、原发性醛固酮增多症、嗜铬细胞瘤及大动脉炎、肾动脉粥样硬化、肾动脉纤维肌性结构不良诊治积累了非常丰富的经验　北京　2013年10月15日

(The diagnosis and treatment of hypertension, very experienced in the diagnosis and treatment of refractory hypertension, primary hyperaldosteronism, pheochromocytoma and Takayasu's arteritis, renal atherosclerosis, and renal artery fibromuscular dysplasia; Beijing; October 15, 2013)

Text 3.冠状动脉支架安置术、瓣膜球囊扩张术、心律失常射频消融术、起搏器安置术，先天性心脏病的封堵术　黑龙江　2012年3月30日

(Coronary stenting, valve balloon dilation, radiofrequency ablation of arrhythmias, pacemaker implantation, closure of congenital heart disease; Heilongjiang; March 30, 2012)

Figure 7 Cardiovascular physician specialty text data example

Analysis of the compiled text data pertaining to physicians' specialties revealed two primary categories of information: diseases (specifically, the doctor's proficiency in diagnosing and treating certain diseases) and treatments (the doctor's expertise in specific treatment methods or approaches). The Unified Medical Language System

(UMLS) offers precise semantic definitions for these information types. In this study, we identified two entity categories within the text data relating to doctor expertise:

Category 1: Diseases. This category typically includes causes leading to a patient's unhealthy state (excluding bad habits) or diagnoses based on a patient's physical condition, which may be treatable or improvable. The corresponding UMLS semantic types include disease or syndrome, injury or poisoning, congenital abnormality, virus or bacterium, molecular or cellular dysfunction, acquired abnormality, malignant progression, cognitive or behavioral impairment, etc.

Category 2: Therapeutic Means (Treatments). This category refers to various treatment procedures or interventions, such as medications, surgical procedures, etc., administered to patients to cure diseases or alleviate symptoms. Relevant UMLS semantic types include pharmacologic substance, therapeutic or preventive procedure, drug delivery device, medical device, steroid, antibiotic, etc.

The method of NER employed in this study transforms the process into a word-based sequence labelling problem. It utilizes CRF in conjunction with the Viterbi algorithm to identify the most probable sequence constituting entities within a given string. Entity sequence labelling in this experiment adheres to the "B, I, O" format, where "B" stands for "beginning" and signifies beginning of a named entity, "I" stands for "inside" and signifies that the word is inside a named entity, "O" stands for "outside" and signifies that the word is just a regular word outside of a named entity. Combining the sequence labelling form with two types of named entities (disease and treatment) yields a total of five label formats, as exemplified in Tab. 2. Three personnel with medical backgrounds were tasked with annotating the training and test datasets. To evaluate annotation consistency, the Intraclass Correlation Coefficient (ICC) was utilized, yielding a value of 0.895. This high ICC score confirms the reliability of the annotation outcomes.

Table 2 Form and meaning of annotation

Named Entity Type	Annotation Form	Annotation Meaning	Annotated Example
Disease	B-Disease	Beginning of Disease Entity	心
	I-Disease	Middle or End of Disease Entity	衰
Treatment	B-Treatment	Beginning of Treatment Entity	搭
	I-Treatment	Middle or End of Treatment Entity	桥
Other	O	Not Part of a Named Entity	等

In this experiment, the model is constructed by the Pytorch deep learning framework and Python programming language. Tab. 3 shows the configuration of the training environment. The BERT model pretrains the language model with the BERT-Base-Chinese model. The model′s foundational architecture consists of 12 layers of stacked bidirectional Transformers, with a hidden layer dimension of 768. It can process text up to a maximum length of 512 characters. In the training process, the fine-tuning learning rate of the pre-trained language model is 3e−5 with the use of Adam optimizer. The dropout rate is set to 0.1 in the input and output layers of BiLSTM to avoid over-fitting problem. Tab. 4 shows the hyperparameter values of the model.

Table 3 Experimental environmentconfiguration

Operating System	Windows 10
CPU	Intel(R) Core(TM) i7-4770HQ 2.20GHz
GPU	Intel(R) Iris(TM) Pro Graphics 5200
Python	3.7.4
Pytorch	1.13.1

Table 4 Model hyperparameter settings

Parameter	Value
Number of Epochs	30
Optimizer	Adam
Batch Size	16
Maximum Input Sequence Length	128
Dropout Rate	0.1
Learning Rate	3e-5
LSTM Hidden Layer Dimension	1.13.1

4.2 Named Entity Recognition Results and Discussion

In this investigation, three metrics were employed to assess the performance of the model: precision, recall, and the $F1$ score. Recall measures the model's capability to detect every entity in the dataset, higher recall means that the model is able to identify a larger proportion of all the true entities. Whereas precision evaluates the model's accuracy in accurately recognizing entities, higher precision indicates that the model's predictions are more accurate, and its reliability is higher. The harmonic mean of precision and recall, $F1$ score, is a metric utilized to assess the overall effectiveness of a model, a higher $F1$ score indicates a better balance between precision and recall. In this experiment, various models were compared to analyze their entity recognition capabilities. The Word2Vec-BiLSTM-CRF model, differing from BERT-BiLSTM-CRF in its word vector training approach but similar in downstream modeling, was utilized. Additionally, BERT-CRF and BERT-BiLSTM models, aligned with BERT-BiLSTM-CRF in the pre-training language model but divergent in the downstream model, were also examined. The objective of this comparative analysis was to assess the entity recognition capabilities of the BERT-BiLSTM-CRF model on the text data utilized in this research. The experiment investigated the BERT model's enhancement effect relative to static word vector representation methods. It also examined the superior performance of the BiLSTM-CRF as a downstream recognition model, highlighting its effectiveness in NER tasks.

The effectiveness of various models in entity recognition was assessed using the dataset from this study, and the outcomes were detailed in Tab. 5. Initially, a comparison was made between the Word2Vec-BiLSTM CRF and BERT-BiLSTM-CRF models in terms of performance. The integration of the BERT pre-training model resulted in diverse enhancements in the metrics of precision, recall, and $F1$ score with respect to entity recognition. Specifically, for disease entities, there was an increase of 1.11% in precision, 3.4% in recall, and 2.26% in the $F1$ score. Furthermore, for treatment entities, there were significant improvements of 42.58% in precision, 20.46% in recall, and 30.71% in the $F1$ score. These outcomes validate the efficacy of BERT in Chinese-named entity recognition tasks within the online medical domain. Subsequently, the performances of the BERT-CRF and BERT-BiLSTM-CRF models were contrasted. For both entity types, the BERT-BiLSTM-CRF model exhibited a notable increase in precision (6.32% and 12.5%), recall (3.36% and 18.18%), and $F1$ score (4.89% and 16.32%). This enhancement is largely attributed to the BiLSTM model's bidirectional structure, which effectively captures contextual sequence information. The final comparison was between the BERT-BiLSTM as well as BERT-BiLSTM-CRF models. Under the background of disease entities, the BERT-BiLSTM-CRF model showed increases of 4.63% in precision, 4.7% in recall, and 4.67% in the $F1$ score. Similarly, for treatment entities, improvements of 11.11% in precision, 9.09% in recall, and 10% in the $F1$ score were observed. These experimental outcomes indicate that the integration of a CRF module can effectively utilize the correlation between adjacent labels. This integration prevents invalid labeling sequences such as "B-Treatment I-Disease ..." and ensures the identification of the globally optimal labeling sequence, thereby enhancing the overall entity recognition performance.

Table 5 Experimental outcomes of named entity recognition with various models

Model	Named Entity Type	Precision / %	Recall / %	$F1$ Score / %
Word2Vec-BiLSTM-CRF	Disease	96.24	95.26	95.74
	Treatment	57.42	61.36	59.29
BERT-CRF	Disease	91.03	95.30	93.11
	Treatment	87.5	63.64	73.68
BERT-BiLSTM	Disease	92.72	93.96	93.33
	Treatment	88.89	72.73	80.00
BERT-BiLSTM-CRF	Disease	97.35	98.66	98.00
	Treatment	100	81.82	90.00

To facilitate a comprehensive analysis of the experimental outcomes, column charts for the disease and treatment entity recognition outcomes of various models, along with a line chart illustrating the $F1$ scores for both entity types, were created based on the data in Tab. 5. These are presented in Fig. 8, 9, and 10. As depicted in Fig. 8 and 9, the BERT-BiLSTM-CRF model outperformed other models in terms of precision, recall, and $F1$ score. A comparison of these two figures reveals a notable disparity in the Word2Vec-BiLSTM-CRF model's recognition effectiveness for disease and treatment entities. Specifically, its performance in recognizing disease entities was significantly better than that for treatment entities. This discrepancy might be attributed to the limitations inherent in static word vectors. Fig. 10 compares the $F1$ scores of various models for the two types of entity recognition. It is observed that the $F1$ score for the two entity types varies significantly across models, with the scores for disease entity recognition consistently surpassing those for treatment entities. This variation could be linked to the structural characteristics of the entities or possibly to the uneven distribution of these entities in the annotated training dataset. Overall, the accuracy of model testing evidently tends to improve with an increase in the number of entities labelled.

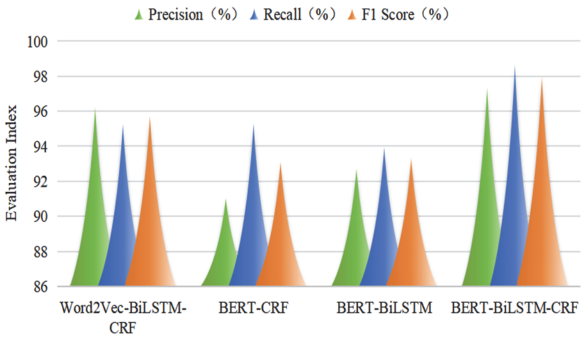


Figure 8 Comparison of disease entity recognition outcomes by various models

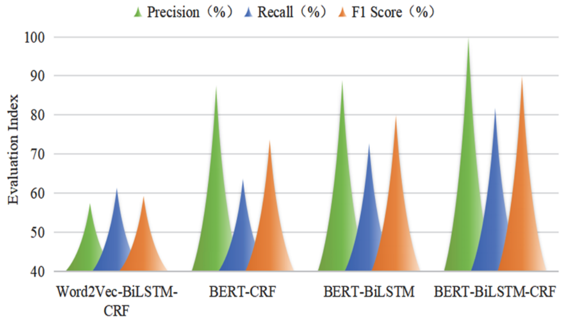


Figure 9 Comparison of treatment entity recognition outcomes by various models

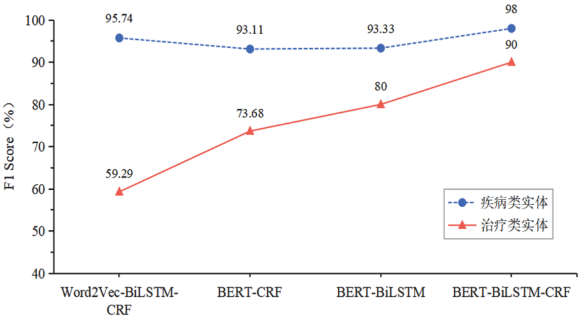


Figure10 Comparison of $F1$ score of various models for two types of entities

5 ANALYSIS

5.1 Analysis of Regional Differences in PhysicianSpecialty

This study delves into the provincial regional disparities and the evolutionary trends in physicians' specialties within online medical communities. This is achieved by utilizing the NER results derived from the BERT-BiLSTM-CRF model. The aim is to extract valuable information from the text data concerning physicians' specialties. These texts predominantly comprise two types of entities: disease and treatment. Variations in doctors' expertise descriptions lead to different instances of these entities. In view of this, the analysis is conducted from both disease and treatment perspectives. This dual approach ensures a comprehensive, effective, and scientific examination of regional differences and evolutionary trends.

The trained model was employed to predict NER for the remaining 9431 text entries, excluding the training and test datasets. Within these, some data were excluded from the analysis due to issues like excessive length or brevity or due to specific description modes (e.g., "心血管临床"),

rendering them ineffective for this study's purposes. Consequently, 9336 valid data entries were obtained for the analysis. Among these, 9164 entries contained recognizable disease-related entities, and 3650 entries included treatment-related entities. These results were statistically aggregated to identify the most frequently occurring entities in both categories, as shown in Tab. 6. Notably, coronary heart disease, hypertension, and arrhythmia emerged as the most prevalent diseases in the cardiovascular field. In terms of treatments, interventional therapy, radiofrequency ablation, pacemaker implantation, and drug therapy were identified as the most common methods in the cardiovascular department.

Table 6 The top 8 disease entities and treatment entities with the highest frequency of occurrence

Disease Entity		Treatment Entity	
Entity	Frequency	Entity	Frequency
Coronary hear disease	7175	Interventional therapy	964
Hypertension	6436	Radio frequency ablation	446
Arrhythmia	4186	Pacemaker implantation	381
Heart failure	3271	Intervene therapy of coronary heart disease	303
Myocardiopathy	2053	Drug therapy	291
Hyperlipidemia	1568	Percutaneous coronary intervention	246
Common diseases in cardiology	1398	Coronary angiography	192
Acute and chronic heart failure	1093	Stent implantation	59

Inverse Document Frequency (*IDF*) is a widely used method in information retrieval to gauge the significance of words [54]. Its fundamental concept involves counting the frequency of a word's appearance across a set of documents. Words appearing in fewer documents are considered to have more discriminatory power. In this study, we adapt the *IDF* algorithm's principle, using the *IDF* value of disease and treatment entities to represent their ability to distinguish doctors' expertise. Specifically, the rarer a disease or treatment entity in physicians' specialty texts, the stronger its capacity to differentiate professional expertise. The calculation of the entity's *IDF* in this study follows the formula in Eq. (15). Here, $|D|$ denotes the total amount of texts in the dataset, and $|D_{entity}|$ refers to the number of texts where the entity appears. Subsequently, we use the calculation method outlined in Eq. (16) to assess the overall treatment proficiency. This proficiency is reflected by physicians' expertise texts in a particular specialty or department within a region. In this equation, *num_largeIDF* is the count of entities in a text with an *IDF* exceeding a set threshold, *num_entity* is the total amount of entities in a text, *text_region* indicates texts relevant to a specific region, and *num_text_region* is the count of related texts in that region. A higher *proportion_region* value implies a greater average proficiency among doctors in that region in treating complex and uncommon diseases and in employing advanced treatment methods. The *proportion_region* for both disease and treatment entities is calculated separately. To visually represent the spatial characteristics of the distribution of cardiovascular physician resources, as indicated by their online medical community expertise texts, we have calculated the *proportion_region* for the expertise level of cardiovascular physicians across different provincial regions. These calculations are visualized spatially in Fig. 11 and 12 for disease and treatment entities, respectively.

$$IDF\_entity = \log\left(\frac{|D|}{1+\left|D_{entity}\right|}\right) \tag{15}$$

$$proportion\_region = \frac{\sum_{text\_region} \dfrac{num\_largeIDF}{num\_entity}}{num\_text\_region} \tag{16}$$

This study utilized provincial data on cardiovascular physician resources in China, excluding Taiwan Province and the two Special Administrative Regions, Hong Kong and Macao, due to a lack of data. A natural discontinuous point grading method was employed to categorize the 31 provinces into five distinct levels. As depicted in Fig. 11 and 12, the spatial distribution of cardiovascular physician resources, as indicated by expertise levels in online medical communities, exhibits noticeable spatial heterogeneity. Different levels, such as high, relatively high, and medium, generally demonstrate regional clustering, with isolated distributions in certain areas. Both Figs identify Beijing, Shanghai, Sichuan, and Jilin as regions with high-level expertise. While the categorization of other regions is not entirely consistent across the two maps, the general trend is similar. Lower-level areas predominantly lie in China's western region; the central region has a slightly higher concentration of higher-level and high-level areas; the eastern region mainly consists of higher-level and high-level areas, with some medium and low-level regions; and the northeast is primarily characterized by higher-level and high-level areas. In summary, there are discernible disparities among the eastern, central, western, and northeastern regions. However, the differences within these regions are considerably less pronounced than those between the western region and the others. The cardiovascular physician resources in the western region, analyzed from the perspective of physician specialty, still trail behind those in the eastern, central, and northeastern regions.

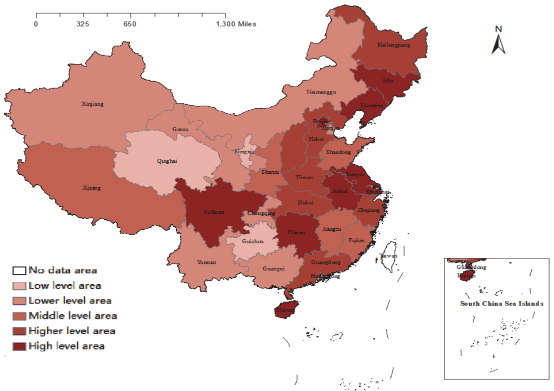


Figure11 Spatial distribution of cardiovascular physicians' expertise level based on disease entity analysis

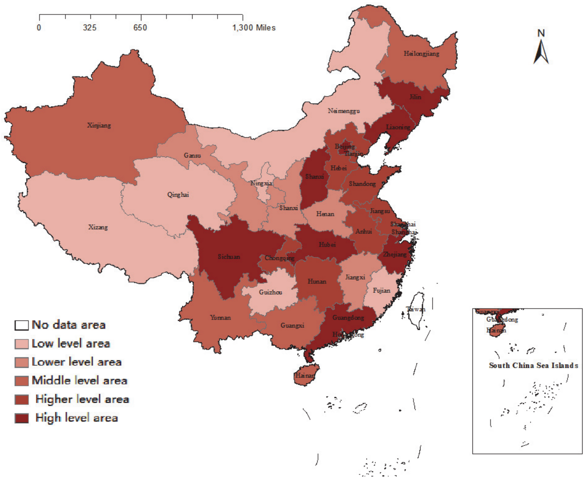


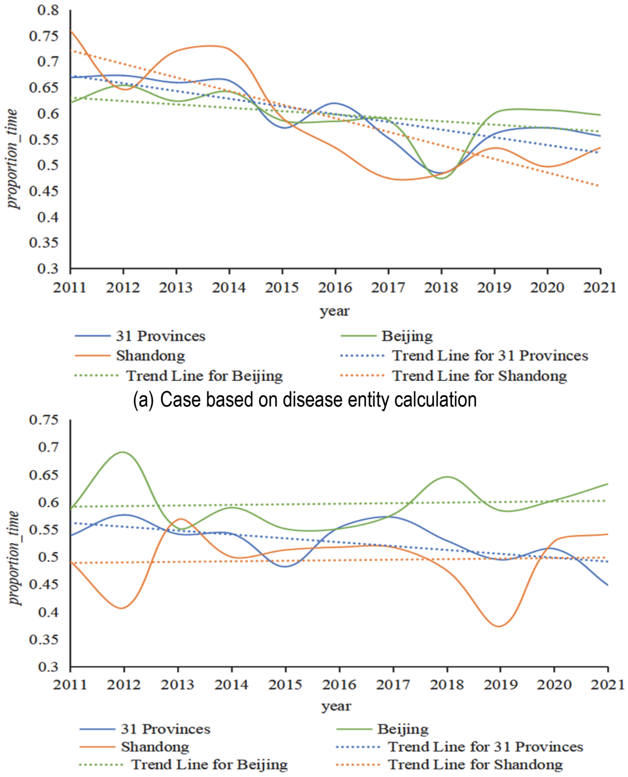
Figure 12 Spatial distribution of cardiovascular physicians' expertise level based on treatment entity analysis

Overall, multiple factors contribute to the lag in physician resources in the western region. First, the economic development level in the western region is low, leading to a lack of well-established medical facilities, slower introduction of advanced medical equipment, limited financial support for medical research and training, and difficulty in attracting talented medical professionals. At the same time, there is significant policy support for the eastern region, with substantial financial and policy subsidies. The gap between the western and eastern regions in this regard is quite evident. Additionally, the complex terrain and inconvenient transportation in the western region may also be factors contributing to the shortage of physician resources.

5.2 Analysis of Evolutionary Trends in Physician Specialty

The preceding section provided a regional difference analysis for the specialty of cardiovascular physicians across 31 provincial-level regions in China, incorporating historical data from various dates corresponding to the creation of physicians' accounts and information on the platform. The subsequent analysis explores the evolutionary trends of cardiovascular physician resources in online medical communities over time, examining the development of doctors' specialties relative to the years when their platform accounts were created. This analysis aims to offer theoretical guidance for online medical community operators to enhance platform development and present a viable method for tracing the evolution of China's physician resources. Continuing with the calculation approach outlined in Eq. (16), the average proficiency level of doctors in managing complex or rare diseases and advanced treatment methods, as reflected in their specialty texts over different time periods, is illustrated as per Eq. (17). In this equation, *num_large IDF* and *num_entity* retain the same meanings as in Eq. (16). *text_time* denotes the texts relevant to a specific time period, while *num_text_time* indicates the number of related texts for that period. A higher *proportion_time* value suggests a greater average mastery level of complex and rare diseases and advanced treatment methods among doctors during that time frame. To provide clear examples, three representative data sets were selected: the entirety of China's 31 provinces, Beijing (noted for its high level of doctor specialty), and Shandong Province (with a substantial volume of doctor specialty data). The evolution of *proportion_time*, calculated based on both disease and treatment entities over time, is demonstrated in Fig. 13a and Fig. 13b respectively.

$$proportion\_time = \frac{\sum\limits_{text\_time} \dfrac{num\_largeIDF}{num\_entity}}{num\_text\_time} \tag{17}$$



(a) Case based on disease entity calculation

(b) Case based on treatment entity calculation

Figure 13 Changes in proportion_time calculated based on two types of entities

For the 31 provinces in China, the analysis of cardiovascular physicians' specialty, based on both disease and treatment entity types, exhibited annual fluctuations. From 2011 to 2021, there was a general downward trend in the measurement level of physicians' specialty, based on both disease and treatment entity analyses. This trend suggests that, for the 31 provinces as a whole, there has been a diminishing differentiation over time in the specialty of doctors who created accounts and provided information on online medical platforms during later stages compared to their earlier counterparts. In Beijing and Shandong Province, the proportion_time calculated based on disease entities also demonstrated a downward trend from 2011 to 2021, with the trend being more pronounced in Shandong than in Beijing. Both regions experienced their most significant decline in 2018, followed by a gradual recovery. Contrasting with the overall declining trend in the 31 provinces, the proportion_time based on treatment entities in Beijing and Shandong exhibited a slight upward trend. Additionally, it was observed that the proportion_time based on treatment entities in Beijing consistently surpassed that of Shandong. This indicates that, compared to Shandong, the differentiation in doctors' specialties, as reflected in the specialty texts of cardiovascular physicians in Beijing, was higher. It also implies that doctors in Beijing, on average, had a higher level of mastery of complex or rare diseases and advanced treatment methods.

6 CONCLUSION AND POLICY IMPLICATIONS

This study, through an analysis of text data concerning doctors' specialties in online medical communities, has developed a BERT-BiLSTM-CRF model based on deep learning techniques. This model efficiently extracts disease and treatment entities from the Chinese texts of doctors' specialties. Experimental results have confirmed the superiority of this model over other NER models. Additionally, the study interprets entity identification results in doctor specialty texts. It introduces methods for measuring the level of doctor specialties across different regions and time periods. Consequently, the study elucidates provincial regional differences and evolutionary trends of cardiovascular physician resources.

Given the extensive presence of doctors on China's "Haodaifu Online" platform, the regional information of doctors in online communities accurately mirrors their offline locations. Therefore, the regional differences in doctors' specialties within these online communities, to a certain extent, reflect the offline regional disparities in relevant specialties. This study concludes that regional imbalances in the distribution of cardiovascular physician resources in China persist. As medical resources are fundamental to public health and physician resources are at the core of these resources, the equitable distribution of physician resources is vital for the sustainable development of the healthcare sector. The uneven distribution of physician resources may be attributed to disparities in economic development, population distribution, culture and education, geography, and transportation conditions across different regions. Despite these challenges, all responsible departments should strive to promote a balanced, coordinated, and comprehensive development of medical and health services. Specifically, the western region of China should acknowledge these regional differences in physician resources, intensify policy support for the medical and health systems, and allocate funds towards the adoption of high-end technology and the development of exceptional talent. This approach could guide the attraction of high-quality doctor resources to the western region. Additionally, in regions where certain specialties lag, precise methodologies should be employed to address these gaps, transitioning from broad health investments to refined resource management rather than solely focusing on expanding resource quantity. For regions with an abundant allocation of physician resources, efforts should be made to prevent resource waste due to redundancy.

The differentiation in the overall expertise of cardiovascular physicians who have recently created accounts and provided information on online medical platforms appears to be diminishing compared to earlier stages. This trend may be attributed to the fact that high-quality doctors either joined the platform or had their information included by platform operators in its early stages. As time progressed, the doctors joining in later stages were predominantly younger and newer to the profession, requiring a period of professional training and skill development. Online medical platforms can strategically recommend these younger doctors to patients based on specific needs, thereby maximizing the potential benefits of new medical talent. Furthermore, for online medical platforms to foster high-quality and positive development, attention must be paid to the decreasing differentiation in the specialties of newer doctors on the platform. Doctors on these platforms play a crucial role in enhancing the quality of online medical care. A doctor's professional level is a primary consideration for patients when selecting a medical provider and is a key component of the competitive edge of major online medical platforms. To attract more patients, platforms need to enlist more doctors skilled in diagnosing and treating difficult or rare diseases and proficient in advanced treatment methods. This strategy would not only expand the platform's scope but also effectively address the medical shortcomings in various specialties.

Contrasting with studies that use panel data to examine the equity of physician resource distribution based on population and geographical area, this research adopts a novel approach by starting from the perspective of physician professional skills. It delves into the differences and evolutionary trends of physician resources across provincial regions in China through in-depth mining of physician specialty text data. While this perspective and method provide valuable insights into the regional disparities and evolutionary trends of doctor resources, optimizing these resources should involve more than just increasing their number. Although this study reveals the spatial disparities in the distribution of physician resources in China, it still has certain limitations, particularly in terms of data coverage and depth of analysis. Notably, Sunet al. (2024) also mentions in his research the significant gap in physician resources between the eastern and western regions, which provides strong support for our analysis [55]. Building upon the findings of these existing studies, we further confirm the impact of factors such as regional economic development and policy support on the distribution of physician resources. However, this study does not take into account other critical factors, such as population and geographical differences, when measuring the level of physician resource allocation, which represents a limitation. Future research could integrate these factors for a more comprehensive analysis. As for the policy recommendations, future studies could delve deeper into how specific policy measures, particularly targeted guidance for the western regions, can help alleviate this imbalance.

Acknowledgements

This work was supported by the National Social Science Fund of China under Grant No. 24CTJ034.

7 REFERENCES

[1] Al-Shboul, M. & Al Rawashdeh, R. (2022). The impact of institutional quality and resources renton health: The case of GCC. *Resources Policy*, *78*, 102804. https://doi.org/10.1016/j.resourpol.2022.102804

[2] Xiong, W., Deng, Y., Yang, Y., Zhang, Y., & Pan, J. (2021). Assessment of medical service pricingin China's health care system: Challenges, constraints, and policy recommendations. *Frontiersin Public Health*, *9*, 787865. https://doi.org/10.3389/fpubh.2021.787865

[3] Chai, K., Zhang, Y., & Chang, K. (2020). Regional disparity of medicalresources anditseffect on mortalityrates in China. *Frontiers in Public Health*, *8*, 8. https://doi.org/10.3389/fpubh.2020.00008

[4] Xu, J., Zheng, J., Xu, L., & Wu, H. (2021). Equity of health services utilisation and expenditure among urban and rural residents under universal health coverage. *International Journal of Environmental Research and Public Health*, *18*(2), 593. https://doi.org/10.3390/ijerph18020593

[5] Byrne, J., Conway, E., McDermott, A. M., Matthews, A., Prihodova, L., Costello, R. W., & Humphries, N. (2021). How the organisation of medical work shapes the everyday work experience sunder pinning doctor migration trends: The case of irish-trained emigrant doctors in Australia. *Health Policy*, *125*(4), 467-473. https://doi.org/10.1016/j.healthpol.2021.01.002

[6] Fu, L., Xu, K., Liu, F., Liang, L., & Wang, Z. (2021). Regional disparity and patients mobility: Benefits and spillover effects of the spatial network structure of the health services in China. *International Journal of Environmental Research and Public Health*, *18*(3), 1096. https://doi.org/10.3390/ijerph18031096

[7] Li, S., Deng, L., Zhang, X., Chen, L., Yang, T., Qi, Y., & Jiang, T. (2022). Deep phenotyping of Chinese electronic health records by recognizing linguistic patterns of phenotypic narratives with a sequence motif discovery tool: Algorithm development and validation. *Journal of Medical Internet Research*, *24*(6), e37213. https://doi.org/10.2196/37213

[8] Sun, Y. L. & Zhang, D. L. (2019). Machine learning techniques for screening and diagnosis of diabetes: A survey. *Tehnički vjesnik*, *26*(3), 872-880. https://doi.org/10.17559/TV-20190421122826

[9] Cai, L., Li, J., Lv, H., Liu, W., Niu, H., & Wang, Z. (2023). Integrating domain knowledge for biomedical text analysis into deep learning: A survey. *Journal of Biomedical Informatics*, *143*, 104418. https://doi.org/10.1016/j.jbi.2023.104418

[10] Chen, W., Qiu, P., & Cauteruccio, F. (2024). MedNER: A service-oriented framework for Chinese medical named-entity recognition with real-world application. *Big Data and Cognitive Computing*, *8*(8), 86. https://doi.org/10.3390/bdcc8080086

[11] Hou, G., Jian, Y., Zhao, Q., Quan, X., & Zhang, H. (2024). Language model based on deep learning network for biomedical named entity recognition. *Methods*, *226*, 71-77. https://doi.org/10.1016/j.ymeth.2024.04.013

[12] Liu, H., Ma, Y., Gao, C., Qi, J., & Zhang, D. (2023). Chinese named entity recognition method for domain-specific text. *Tehnički vjesnik*, *30*(6), 1799-1808. https://doi.org/10.17559/TV-20230324000477

[13] Chen, Z., Song, Q., Wang, A., Xie, D., & Qi, H. (2022). Study on the relationships between doctor characteristics and online consultation volume in the online medical community. *Healthcare, 10*(8), 1551. https://doi.org/10.3390/healthcare10081551

[14] Zhang, Y., Strauss, J., & Liu, L. (2021). An OLS and GMM combined algorithm for text analysis for heterogeneous impact in online health communities. *Tehnički vjesnik*, *28*(2), 587-597. https://doi.org/10.17559/TV-20210121100916

[15] Li, M., Bi, X., Wang, L., Han, X., Wang, L., &Zhou, W. (2022). Text similarity measurement method and application of online medical community based on density peakclustering. *Journal of Organizational and End User Computing*, *34*(2), 1-25. https://doi.org/10.4018/JOEUC.302893

[16] Peng, J., Fu, C., Guo, Y., & Huang, L. (2025). The impact of customer value co-creation in online medical consultation on customer service well-being. *Journal of Research in Interactive Marketing*.

[17] Fan, J., Geng, H., Liu, X., & Wang, J. (2022). The effects of online text comments on patients'choices: The mediating roles of comment sentiment and comment content. *Frontiers in Psychology*, *13*, 886077. https://doi.org/10.3389/fpsyg.2022.886077

[18] Ma, C. (2021). Will online medical community participation affect physicians' offline service volume and their diagnosis and treatment revenue: an empirical study based on the PSM-DID model. *Chinese Journal Health Policy*, *14*(9), 47-53.

[19] Su, Y., Wang, M., Wang, P., Zheng, C., Liu, Y., & Zeng, X. (2022). Deep learning joint models forex tracting entities and relations in biomedical: a survey and comparison. *Briefings in Bioinformatics*, *23*(6), bbac342. https://doi.org/10.1093/bib/bbac342

[20] Zhen, Y., Li, Y., Zhang, P., Yang, Z., & Zhao, R. (2023). Frequent words and syntactic context integrated biomedical discontinuous named entity recognition method. *The Journal of Super computing*, *79*, 13670-13695. https://doi.org/10.1007/s11227-023-05224-0

[21] Chen, W., Qiu, P., & Cauteruccio, F. (2024). MedNER: A service-oriented framework for Chinese medical named-entity recognition with real-world application. *Big Data and Cognitive Computing*, *8*(8), 86. https://doi.org/10.3390/bdcc8080086

[22] Bhatia, S., Alojail, M., Sengan, S., & Dadheech, P. (2022). An efficient modular framework for automatic LIONC classification of MedIMG using unified medical language. *Frontiers in Public Health*, *10*, 926229-926229. https://doi.org/10.3389/fpubh.2022.926229

[23] Lee, L. & Lu, Y. (2021). Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2801-2810. https://doi.org/10.1109/JBHI.2020.3048700

[24] Cho, M., Ha, J., Park, C., & Park, S. (2020). Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*, *103*, 103381. https://doi.org/10.1016/j.jbi.2020.103381

[25] Wang, C., Wang, H., Zhuang, H., Li, W., Han, S., Zhang, H., & Zhuang, L. (2020). Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. *Journal of Biomedical Informatics*, *111*, 103583. https://doi.org/10.1016/j.jbi.2020.103583

[26] Yin, M., Mou, C., Xiong, K., & Ren, J. (2019). Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *Journal of Biomedical Informatics*, *98*, 103289. https://doi.org/10.1016/j.jbi.2019.103289

[27] An, Y., Xia, X., Chen, X., Wu, F. X., & Wang, J. (2022). Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artificial Intelligence in Medicine*, *127*, 102282. https://doi.org/10.1016/j.artmed.2022.102282

[28] Li, Y., Du, G., Xiang, Y., Li, S., Ma, L., Shao, D., Wang, X., & Chen, H. (2020). Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge. *Journal of Biomedical Informatics*, *106*, 103435. https://doi.org/10.1016/j.jbi.2020.103435

[29] Liu, N., Hu, Q., Xu, H., Xu, X., & Chen, M. (2021). Med-BERT: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, *18*(8), 5600-5608. https://doi.org/10.1109/TII.2021.3131180

[30] Shi, X., Yi, Y., Xiong, Y., Tang, B., Chen, Q., Wang, X., Ji, Z., Zhang, Y., & Xu, H. (2019). Extracting entities with

attributes in clinical text via joint deep learning. *Journal of the American Medical Informatics Association*, *26*(12), 1584-1591. https://doi.org/10.1093/jamia/ocz158

[31] Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M., & Xu, H. (2014). A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, *21*(5), 808-814. https://doi.org/10.1136/amiajnl-2013-002381

[32] Zhao, S., Cai, Z., Chen, H., Wang, Y., Liu, F., & Liu, A. (2019). Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *Journal of Biomedical Informatics*, *99*, 103290. https://doi.org/10.1016/j.jbi.2019.103290

[33] Qin, Q., Zhao, S., & Liu, C. (2021). A BERT‐BiGRU‐CRFmodel for entity recognition of Chinese electronic medical records. *Complexity*, 2021, 6631837. https://doi.org/10.1155/2021/6631837

[34] Zhao, Z., Yang, S., Zhao, Y., Chen, H., Dou, N., He, G., Sun, Z., Yang, Y., Luo, J., Gao, H., Dai, S., & Chen, C. (2021). Status quo and equity analysis of human resources for health in China: based on five-year data. *Journal of Chinese Human Resources Management*, *12*(1), 77-85. https://doi.org/10.47297/wspchrmWSP2040-800506.20211201

[35] Yu, H., Yu, S., He, D., & Lu, Y. (2021). Equity analysis of Chinese physician allocation based on Gini coefficient and Theil index. *BMC health services research*, *21*(1), 455. https://doi.org/10.1186/s12913-021-06348-w

[36] Xiong, X., Jin, C., Chen, H., & Luo, L. (2016). Using the fusion proximal area method and gravity method to identify areas with physician shortages. *PLoS One*, *11*(10), e0163504. https://doi.org/10.1371/journal.pone.0163504

[37] Erdenee, O., Paramita, S. A., Yamazaki, C., & Koyama, H. (2017). Distribution of health care resources in Mongolia using the Gini coefficient. *Human Resources for Health*, *15*(1), 56. https://doi.org/10.1186/s12960-017-0232-1

[38] Paramita, S. A., Yamazaki, C., Setiawati, E. P., & Koyama, H. (2018). Distribution trends of Indonesia's health care resources in the decentralization era. *The International Journal of Health Planning and Management*, *33*(2), e586-e596. https://doi.org/10.1002/hpm.2506

[39] Wang, S., Xu, J., Jiang, X., Li, C., Li, H., Song, S., Huang, E., & Meng, Q. (2018). Trends in health resource disparities in primary health care institutions in Liaoning Province in Northeast China. *International Journal for Equity in Health*, *17*(1), 178. https://doi.org/10.1186/s12939-018-0896-8

[40] Pál, V., Lados, G., Makra, Z. I., Boros, L., Uzzoli, A., & Fabula, S. Concentration and inequality in the geographic distribution of physicians in the European Union. *Regional Statistics*, *11*(3), 3-28. https://doi.org/10.15196/RS110308

[41] Yan, X., He, S., Webster, C., & Yu, M. (2022). Divergent distributions of physicians and healthcare beds in China: Changing patterns, driving forces, and policy implications. *Applied Geography*, *138*, 102626. https://doi.org/10.1016/j.apgeog.2021.102626

[42] Wang, L., Hu, Z., Chen, H., Zhou, C., Tang, M., & Hu, X. (2024). Differences in regional distribution and inequality in health workforce allocation in hospitals and primary health centers in China: A longitudinal study. *International Journal of Nursing Studies*, *157*, 104816. https://doi.org/10.1016/j.ijnurstu.2024.104816

[43] Montañez-Hernández, J. C., Alcalde-Rabanal, J., & Reyes-Morales, H. (2020). Socioeconomic factors and inequality in the distribution of physicians and nurses in Mexico. *Revista de Saúde Pública*, *54*, 58. https://doi.org/10.11606/s1518-8787.2020054002011

[44] Ma, C., Ge, J., & Han, Y. (2024). Global rounds: Advancing cardiovascular health in China. *Circulation*, *151*(6), 340-342. https://doi.org/10.1161/CIRCULATIONAHA.124.071544

[45] Chen, P., Zhang, M., Yu, X., & Li, S. (2022). Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT.*BMC Medical Informatics and Decision Making*, *22*(1), 315. https://doi.org/10.1186/s12911-022-02059-2

[46] Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., & Kang, J. (2019). A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, *7*, 73729-73740. https://doi.org/10.1109/ACCESS.2019.2920708

[47] Wen, G., Chen, H., Li, H., Hu, Y., Li, Y., & Wang, C. (2020). Cross domains adversarial learning for Chinese named entity recognition for online medical consultation. *Journal of Biomedical Informatics*, *112*, 103608. https://doi.org/10.1016/j.jbi.2020.103608

[48] Shi, J., Sun, M., Sun, Z., Li, M., Gu, Y., & Zhang, W. (2022). Multi-level semantic fusion network for Chinese medical named entity recognition. *Journal of Biomedical Informatics*, *133*, 104144. https://doi.org/10.1016/j.jbi.2022.104144

[49] Gong, L., Zhang, Z., & Chen, S. (2020). Clinical named entity recognition from Chinese electronic medical records based on deep learning pretraining. *Journal of Healthcare Engineering*, *2020*(1), 8829219. https://doi.org/10.1155/2020/8829219

[50] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[51] Tingjiang, T., Enyuan, W., Ke, Z., & Changfang, G. (2023). Research on assisting coal mine hazard investigation for accident prevention through text mining and deep learning. *Resources Policy*, *85*, 103802. https://doi.org/10.1016/j.resourpol.2023.103802

[52] Yu, W., Gonzalez, J., & Li, X. (2021). Fast training of deep LSTM networks with guaranteed stability for nonlinear system modeling. *Neurocomputing*, *422*, 85-94. https://doi.org/10.1016/j.neucom.2020.09.030

[53] Tan, K. S., Lim, K. M., Lee, C. P., & Kwek, L. C. (2022). Bidirectional long short-term memory with temporal dense sampling for human action recognition. *Expert Systems with Applications*, *210*, 118484. https://doi.org/10.1016/j.eswa.2022.118484

[54] Juraev, G. & Bozorov, O. (2023). Using TF-IDF in text classification. *AIP Conference Proceedings*, *2789*(1), 050017. https://doi.org/10.1063/5.0145520

[55] Sun, Y. & Wu, S. (2024). Study on comparison of health resource allocation efficiency and influence path in eastern central and western regions of China. *Medicine and Society*, *37*(4), 61-67.

Contact information:

Siyu WANG
School of Information Science
Guangdong University of Finance and Economics, China
No. 21 Xueyuan Road, Baiyun District, Guangzhou, Guangdong Province, China
E-mail: wangsiyuletter@163.com

Mingyang LI
(Corresponding author)
School of Economics and Management
Changchun University of Technology,
China
No. 1699 Qingshan Road, Chaoyang District, Changchun, Jilin Province, China
E-mail: mingyang_1208@163.com

Limin WANG
(Corresponding author)
School of Information Science
Guangdong University of Finance and Economics, China
No. 21 Xueyuan Road, Baiyun District, Guangzhou, Guangdong Province, China
E-mail: wlm_new@163.com

Xuming HAN
College of Information Science and Technology
Jinan University,
China
No. 601, Huangpu Avenue West, Tianhe District, Guangzhou, Guangdong
Province, China
E-mail: hanxuming@jnu.edu.cn

Jiawei LI
School of Economics and Management
Changchun University of Technology, China
No. 1699 Qingshan Road, Chaoyang District, Changchun, Jilin Province, China
E-mail: lijiawei@ccut.edu.cn