Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.036831
Article

Tech Science Press

Data Masking for Chinese Electronic Medical Records with Named Entity Recognition

Tianyu He[1], Xiaolong Xu[1,*], Zhichen Hu[1], Qingzhan Zhao[2], Jianguo Dai[2] and Fei Dai[3]

[1]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 21000, China
[2]Geospatial Information Engineering Research Center, Xinjiang Production and Construction Corps, Shihezi, 832003, China
[3]College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, 650224, China
*Corresponding Author: Xiaolong Xu. Email: xlxu@ieee.org
Received: 13 October 2022; Accepted: 13 December 2022

Abstract: With the rapid development of information technology, the electronification of medical records has gradually become a trend. In China, the population base is huge and the supporting medical institutions are numerous, so this reality drives the conversion of paper medical records to electronic medical records. Electronic medical records are the basis for establishing a smart hospital and an important guarantee for achieving medical intelligence, and the massive amount of electronic medical record data is also an important data set for conducting research in the medical field. However, electronic medical records contain a large amount of private patient information, which must be desensitized before they are used as open resources. Therefore, to solve the above problems, data masking for Chinese electronic medical records with named entity recognition is proposed in this paper. Firstly, the text is vectorized to satisfy the required format of the model input. Secondly, since the input sentences may have a long or short length and the relationship between sentences in context is not negligible. To this end, a neural network model for named entity recognition based on bidirectional long short-term memory (BiLSTM) with conditional random fields (CRF) is constructed. Finally, the data masking operation is performed based on the named entity recognition results, mainly using regular expression filtering encryption and principal component analysis (PCA) word vector compression and replacement. In addition, comparison experiments with the hidden markov model (HMM) model, LSTM-CRF model, and BiLSTM model are conducted in this paper. The experimental results show that the method used in this paper achieves 92.72% Accuracy, 92.30% Recall, and 92.51% F1_score, which has higher accuracy compared with other models.

Keywords: Named entity recognition; Chinese electronic medical records; data masking; principal component analysis; regular expression

1 Introduction

China has a large population and well-supported medical institutions, which makes the volume of medical record data huge [1]. And with the rapid development of big data technology in recent years, electronic medical records have gradually become a trend to replace traditional paper medical records. Compared with traditional paper medical records, electronic medical records have the advantages of easy long-term preservation and data retrieval, and reduce the waste of forest resources needed to produce paper [2]. The application of electronic medical records is an important part of the realization of medical intelligence. The National Health and Wellness Commission has proposed that the country attached great importance to medical big data services. Our country is preparing to build a national unified management system to realize the interoperability of big data of electronic medical records. However, electronic medical records contain a large amount of personal privacy information, and how to achieve privacy protection is an issue that cannot be ignored. To prevent information leakage and theft, the National Health Planning Commission is strictly supervising the management of medical record data by medical institutions. And this paper takes this as the starting point to propose data masking for Chinese electronic medical records with named entity recognition.

To extract key information from the huge amount of Chinese electronic medical records data, information extraction (IE) is needed. IE consists of three parts: named entity recognition (NER), relation extraction and event extraction. Named entity recognition is in continuous development, from the initial rule-based methods to the previous methods based on traditional machine learning such as HMM [3] and CRF [4], and then to the current methods based on deep learning [5]. Before entering the named entity recognition model, it is necessary to transform the text data into the form of word vectors required for the model input. NER usually consists of two steps, the first step is to delineate the boundary and separate each word for recognition, and the second step is to identify the type of entity. Compared with English NER, Chinese NER is more difficult to implement. The main reason is that Chinese and English have different writing styles and word separations [6]. Chinese named entity recognition technique used in this paper extracts entities related to people's names, organizations, and locations from text data and labels their entity types. The implementation of named entity recognition provides the next step for data masking of private entities.

Data privacy protection cannot be ignored in the big data environment [7]. Electronic medical records data in the medical field is gradually becoming the hardest hit area for privacy leakage because it contains a huge amount of patient privacy [8]. Therefore, privacy protection needs to be realized by using technical means. Data masking [9] can be used to solve this problem. The sensitive data contained in the original data is encrypted and replaced successively to reduce the risk of data leakage [10]. There are also certain problems with data masking, such as incomplete data masking, loss of valuable information in data, and inconsistent data masking standards. In this paper, we propose two methods for data masking of electronic medical record data due to the different types of privacy entities. The first one is to use regular expressions combined with a hash encryption [11] algorithm to achieve masking for the numeric type of private entity data. The second one is to use PCA [12] word vector compression followed by word vector replacement operation to desensitize textual privacy entity data.

The main contributions of this paper are as follows:

1) In this paper, the BiLSTM-CRF model is applied to the privacy data masking and encryption of Chinese electronic medical records by combining regular expressions and PCA compressed word vector algorithm.
2) For textual entities, the People's Daily corpus is first trained to generate high-dimensional word vectors. Then the word vector is compressed to one dimension by PCA and saved as a word

vector dictionary. Finally, word vector replacement is used to realize data masking. For numeric entities, data masking is implemented using regular expression filtering followed by hash encryption.

3) The results show that the method proposed in this paper effectively improves the named entity recognition performance of Chinese electronic medical record privacy entities, which is significantly better than the classical models such as HMM, LSTM-CRF, and BiLSTM.

The remainder of this paper is organized as follows. Section 2 introduces the named entity recognition of medical privacy data and the related research results and status of the model. Section 3 presents the implementation of BiLSTM-CRF-based named entity recognition. Section 4 presents the implementation of privacy entity data masking. Section 5 is the comparison and analysis of the experimental results. Section 6 is a summary reflection of the whole paper.

2 Related Work

2.1 Rule-Based NER

Rule-based named entity recognition is the original entity recognition method [13], which is mainly applied to entities with contextual links or entities with special formats. Considering the cost and effectiveness of named entity recognition in the judicial field, Jiao et al. [14] proposed a rule-based regular expression method for entity recognition to achieve the task of judicial language entity extraction. Cenikj et al. [15] adopted a rule-based named entity recognition approach that enables the extraction of soft and technical skills using textual data such as job postings, resumes, shouts, performance evaluations, etc. Khaing et al. [16] used a rule-based named entity recognition method in the stock domain to extract the methods and trends. Most such rule-based NERs have laws to explore or build knowledge of a particular specialty into a dictionary of knowledge. Therefore, these methods have shortcomings in improving the accuracy of recognition.

2.2 Machine Learning-Based NER

Machine learning-based NER is also feature-based NER, which has lots of models to choose from, such as CRF models, HMM models, decision tree models [17,18], etc. NER based on traditional machine learning is to study the task as a sequence labeling problem, in which case, it is necessary for the model to be able to relate both the front and back labels of the sequence in the prediction process, and to focus on the relevance, linkage, and dependency of the labels. As early as 1991, HMM was used for the NER task [19]. Li et al. gave full play to the advantages of machine learning based on electronic medical record dataset with self-annotated rules. And obtained good experimental results using the CRF model, which was better than the evaluation indexes of existing studies [20]. Yi et al. [21] applied the CRF model to security entity recognition, combined with regular expression and entity dictionary, and the experimental effect is better than the traditional rule-based model. Gupta et al. [22] used machine learning-based random forest to predict the categories of confirmed, dead, and cured cases of neocoronavirus pneumonia in India. Although the introduction of traditional machine learning methods into NER has improved certain efficiency, traditional machine learning still has many drawbacks. For example, it requires a large amount of manually labeled corpus as the training set, which is time-consuming and laborious.

2.3 Deep Learning-Based NER

With the rapid development and maturity of deep learning in recent years, adding deep learning to NER has also become an implementation method. The main reason is that deep learning will save time and effort and reduce the tedium of manually constructing features. Moreover, deep learning makes gradient descent update models better and the NER task is suitable for the process of nonlinear transformation. So, through the rapid development of recent years, a large number of experts and scholars are applying

BiLSTM-CRF to various fields. Hu et al. applied BiLSTM-CRF to the extraction of government social media comments, and the F1_score of the experiment reached 84.01% [23]. Wang also added an attention mechanism to BiLSTM-CRF for named entity recognition in the Chinese hypertension treatment literature and obtained an F1_score of 86.2% [24]. Hu et al. [25] combined BiLSTM-CRF in the field of sedimentology to identify specific sentence components to achieve sedimentological information extraction. Zhang et al. [26] proposed BiLSTM-CRF-based cross-domain migration to improve Chinese clinical named entity recognition for example (e.g., disease, symptom, drug, anatomy), and obtained an F1_score of 85.43% in comparison with other models. Duppati et al. proposed a deep learning-based entity recognition method for the Indian language in English based on convolutional neural network (CNN), Bi-LSTM, and CRF, which achieved better scores in experimental comparison [27]. Wang et al. conducted an in-depth study on the field of online consultation and validated the effectiveness of the BiLSTM-CRF model for diabetic entity recognition [28].

In this paper, the BiLSTM-CRF model is applied to the recognition of named entities in Chinese electronic medical records. The model is trained with a rich corpus as the training set to optimize the parameters. And the entity recognition effect is satisfactory, which provides the accuracy guarantee for the subsequent entity data masking.

3 Implementation of Named Entity Recognition for Sensitive Data Based on BiLSTM-CRF Model

In this section, the implementation of Chinese named entity recognition based on the BiLSTM-CRF model is introduced. First, the training dataset needs to be pre-processed [29] to complete the work on lexical annotation conversion. Then is the work of converting the word vectors into the input form required by the model. Finally, the model is built and trained to achieve the function of named entity recognition. The overall workflow is shown in Fig. 1:

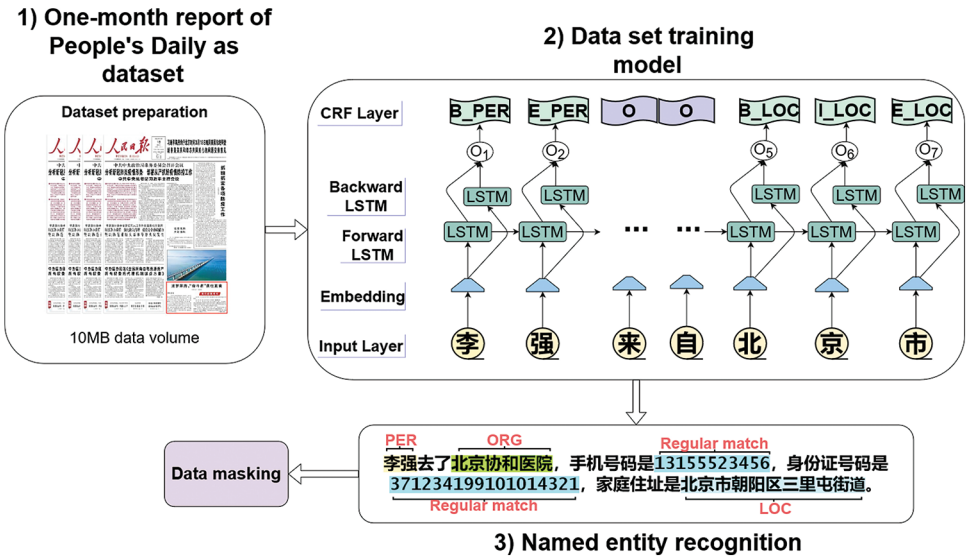

Figure 1: Named entity recognition overall workflow

3.1 Dataset Annotation Conversion

The training dataset used in this paper is a collection of reports related to the People's Daily, which was processed and completed by the Peking University Research Institute in 1998. This dataset has been tagged with words such as "新华社/nt" , "北京/ns", and "李明/nr". In this paper, we focus on the lexical annotation

of three types of entities in the dataset, such as "nr", "ns", and "nt". Then, lexical annotation conversion and sequence annotation work need to be performed. The conversion work has the following steps. Firstly, "nr", "nt", and "nr" are converted to person (PER), organization (ORG), and location (LOC) respectively. Next, the sequence annotation method of BIEO [30] was adopted in this paper. Under the BIEO annotation method, each character in the dataset needs to be annotated. The B (Begin) tag indicates the start character of the entity, the I (Inner) tag indicates the middle character of the information entity, the E (End) tag indicates the end character of the information entity, and the O (Other) tag indicates a character that is not used as the information entity. Taking the label "nr" as an example. First, read the lexically labeled data in turn and record it as *T*. Judge the length of *T*. If the length of *T* is *n* ($n >= 1$), the first word is labeled "*T* [1] B-PER", the last word is labeled "*T*[*n*] E-PER", and all intermediate words are labeled "*T* [2] I-PER, …, $T[n-1]$ I-PER". The other lexical categories are also converted using this method.

3.2 Named Entity Recognition for Chinese Electronic Medical Records

3.2.1 BiLSTM-Based Prediction of Sequence Labels

Since Recurrent Neural Network (RNN) can only handle short-term dependencies, if the processed sentence sequences are too long, the network has no memory function for the previous sentence sequences, which may lead to gradient disappearance. Therefore, to solve this problem, LSTM networks, which are improved based on RNNs, are proposed [31]. The BiLSTM used in this paper, on the other hand, is a combination of a forward LSTM and a backward LSTM network.

The LSTM is controlled by three main gates, namely the input gate, the forgetting gate, and the output gate. It is the combination of these structures that gives the LSTM the function of long-time memory. The specific implementation of LSTM is formulated as follows:

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

$$i_t = \sigma(W_i\ [h_t - 1,\ x_t] + b_i) \tag{2}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_t - 1,\ x_t] + b_c) \tag{3}$$

$$C_t = f_t * c_t - 1 + i_t * \widetilde{C}_t \tag{4}$$

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

$$h_t = o_t * \tanh(C_t) \tag{6}$$

where σ represents the sigmoid activation function, which reduces the result to a coefficient between 0 and 1. b_f, b_i, and b_o represent the bias terms. *tanh* is the hyperbolic tangent activation function. W is the weighting matrix. f_t, i_t, and o_t represent the forgetting gate, the input gate, and the output gate. h_t is the output of the network and *C* represents the memory cell.

Therefore, to enable LSTM networks to make better use of contextual information, the BiLSTM network is introduced, where the BiLSTM includes the forward LSTM and the backward LSTM [32]. For the input sentence, the implicit sequence $\left(\overrightarrow{h_{R0}},\ \overrightarrow{h_{R1}},\ \ldots,\ \overrightarrow{h_{Rn}}\right)$ will be obtained after the forward LSTM, and the sequence $\left(\overleftarrow{h_{L0}},\ \overleftarrow{h_{L1}},\ \ldots,\ \overleftarrow{h_{Ln}}\right)$ will be obtained after the backward LSTM. Finally, the forward and backward implicit state sequence will be spliced to get a whole hidden state sequence $(h_0,\ h_1,\ \ldots,\ h_n)$.

3.2.2 BiLSTM Combined with CRF

After processing by the previous BiLSTM layer calculation, the predicted label corresponding to each word will be obtained, but they do not fully consider the relationship between contextual labels and have

great limitations. However, in the privacy information entity recognition task, the labels of adjacent characters are linked in a backward and forward order, and the strong correlation of labels cannot be neglected, so CRF is used to solve this problem [33]. And the CRF layer is also able to add some constraints to make the prediction results more accurate. For example, the entity tag needs to start with "B-" or "O-" instead of "I-". The combination of entity tag like "B-PER, I-PER, E-PER" is an entity but like "B-PER, E-LOC" format is wrong. Therefore, it is significant to introduce CRF into our privacy entity identification model. The CRF prediction score is achieved by Eq. (7):

$$s(X,\ y) = \sum_{i=0}^{n} Ay_i,\ y_i + 1 + \sum_{i=1}^{n} P_i,\ y_i \tag{7}$$

where $X = \{x_{1,}x_2x_3,\ \ldots,\ x_{n,}\}$ represents the input sequence of sentences. P is the matrix output from the BiLSTM layer. $P_{i,j}$ represents the probability of the j-th tag of the i-th character in the sentence. A is the conversion matrix, $A_{i,j}$ denotes the probability of transferring label i to label j. Finally, for the input sequence X, the formula for the probability of all possible label sequences is defined as:

$$p(y|X) = \frac{e^{S(X,y)}}{\sum_{\bar{y}\in YX} e^{S(X,\bar{y})}} \tag{8}$$

Then, logarithmic operations need to be performed with the help of the log-likelihood function:

$$log(p(y|X)) = s(X,\ y) - log\left(\sum_{\bar{y}\in YX} e^{s(X,\bar{y})}\right) \tag{9}$$

Finally, the most correct prediction result y^* is obtained as:

$$y^* = \underset{\bar{y}\in YX}{argmax}\, s(X,\ \bar{y}) \tag{10}$$

4 Data Masking Implementation for Privacy Entities

In this section, we desensitize the data of the private entities identified in the previous section. Two main approaches are used to achieve data masking. The first one is to use the regular expression [34] to filter digital entities and combine them with a hash encryption algorithm to achieve data masking. The second one is to use the PCA word vector compression algorithm [35] to train the dictionary and then compress it into one-dimensional word vectors. The data masking function of text entities is achieved by word vector replacement. The overall process of data masking is shown in Fig. 2:

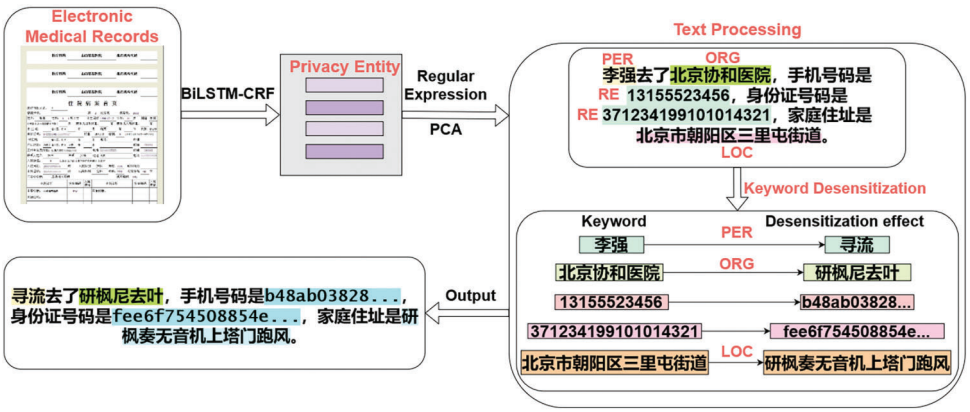

Figure 2: The overall process of data masking

4.1 Regular Expression Filtering

The regular expression is an indispensable technique in the field of natural language processing. It is very effective for filtering and modifying specific content. And it is a sequence of designed characters and character combinations that can be used as a "rule" to filter the desired data.

Algorithm 1: Regular Matching

Input: *Str, Str_re*;

Output: Matching Result *r*;

```
1 Set str_l ← Str.length;
2 Set str_re_l ← Str_re.length;
3 for i to range (str_l) do
4       if str_l [i] is equal to str_re_l [0]
5             j ← i;
6             for h to range (str_re_l) do
7                   if Str_re [h] is equal to Str [j]
8                         Str [j] append to r;
9                   else:
10                        break;
11                  end if
12            end for
13      end if
14 end for
15 Return r;
```

Algorithm 1 describes the algorithm for regular expressions. First, get the length of the string Str and the regular expression *Str_re* respectively (lines 1 and 2). Next, if a character of *Str* matches the first character of *Str_re*, the position index of the character in *Str* is recorded. Then it enters the inner regular expression loop and saves the corresponding character whenever the match is successful. If one match fails, the inner loop (lines 3 to 14) is exited. Finally, the matched string is returned.

For digital privacy entities with entity types 'ID number' and 'cell phone number', specific matching rules to filter all the data that meet the requirements to the maximum extent are designed in this paper.

For the regular filtering of 'ID number', special rules from the actual ID number format of Chinese people are designed in this paper. We need to consider three cases, the first is the beginning of the ID card number, for different provinces in China, the first two codes of ID cards are different, divided into eight forms with the numbers 1–8 as the beginning. The second is the number of bits of the ID card number, the total length of China's resident ID card is 18 bits. The third is the end of the ID card number, which is divided into two forms: pure numbers and the letter X.

Based on this, the matching rules for 'ID number' type entities in this paper are designed as a style, for example, $(\text{^}1[1,\ 2,\ 3,\ 4,\ 5]\backslash d\{13,\ 17\}x\$)$ or $(\text{^}[2, 3, 4, 5, 6]\backslash d\{15, 19\})$. The regular expression filtering above basically includes the format of ID cards for each province in China, and there is a certain amount of fault tolerance. For example, the initial ID digit is overwritten by two digits or underwritten by two

digits, or in some cases, the last digit of the ID number was originally an uppercase X but was written as a lowercase x. All these cases will be filtered out to ensure that the data is completely filtered and extracted.

The regular filtering of "cell phone numbers" is based on the number segments of the three major Chinese telecom operators. The mainstream has six forms of numbering at the beginning, and the length of the phone number is a fixed 11-digit pure number. Based on this, the matching rules for "cell phone number" type entities in this paper are designed to be styled as, for example, (^13\d{7, 11}) or (^14[5|7]\d{6, 10}). The regular expression rules for cell phone numbers are also designed to be fault-tolerant. For example, two extra or two less cell phone numbers are allowed due to initial input errors, which covers most Chinese cell phone number formats.

Then for the filtered privacy entities, they need to be encrypted, and hash encryption will be used as a method of privacy encryption because of its irreversibility. So, with the help of the secure hash algorithm (SHA-256) [36] for hash encryption. The implementation of the SHA-256 algorithm goes through constant initialization (calculation of 8 hash primaries, 64 hash constants), message preprocessing (padding bits, additional length), and logical operations. And finally calculates the message digest to obtain an encrypted string of 64 bits in length in hexadecimal. Then a truncation operation is performed on top of this to keep only the middle 18 characters of the encrypted string.

4.2 PCA-Based Word Vector Compression Encryption Algorithm

For the generated word vectors, the paper introduces the PCA algorithm to compress and reduce the dimensionality of the word vectors [37]. As one of the classic machine learning algorithms, PCA is often used for data compression and feature extraction operations because it is very effective in data dimensionality reduction. The central idea of PCA is to change the dimensionality of the data. For example, compressing m-dimensional data to n-dimensions and preserving the integrity of the data as much as possible. The main principle of PCA is the need to first find a direction on the coordinate axis. Then the projection of the data on the axis is made most discrete and the variance of the data on the axis needs to be maximized. The coordinate axes are orthogonal to each other, and the direction of the next coordinate axis is the direction with the second largest variance. Then keep looping this process, the number of loops is the number of data dimensions at the very beginning. After the previous loop, most of the variance is contained in some of the initial axes, and the end axes contain almost no variance, which allows us to ignore it to realize the dimensionality reduction of the data.

The mathematical formulas used in the PCA algorithm such as variance, covariance, and covariance matrix are implemented as follows:

$$var(x) = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2}{m} \tag{11}$$

$$cov(x,\ y) = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{m} \tag{12}$$

$$C = \begin{pmatrix} cov(x,\ x) & cov(x,\ y) \\ cov(y,\ x) & cov(y,\ y) \end{pmatrix} = \begin{pmatrix} \frac{1}{m}\sum_i^m x_i^2 & \frac{1}{m}\sum_i^m x_i y_i \\ \frac{1}{m}\sum_i^m y_i x_i & \frac{1}{m}\sum_i^m y_i^2 \end{pmatrix} \tag{13}$$

where x_i and $y_i$are the column vectors, $\bar{x}$ and $\bar{y}$ are the mean values of the column vectors, and m is the number of column vectors.

The implementation of the PCA algorithm to compress word vectors is shown in Algorithm 2.

Algorithm 2: PCA

Input: Sample matrix D [], N;

Output: Reduced dimensional matrix O [];

1 $\bar{x} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$;

2 $W \leftarrow$ Each element in $D - \bar{x}$;

3 $Cov\_matrix \leftarrow \frac{1}{m-1} W^T W$;

4 Calculate the eigenvalues and eigenvectors of Cov_matric, and sort eigenvalues from largest to smallest;

5 The eigenvectors corresponding to the first N eigenvalues are combined to form the new matrix F;

6 $O \leftarrow$ Multiply the matrix N with the matrix F;

7 Return O;

Algorithm 2 describes the implementation of PCA. First, calculate the matrix mean $\bar{x}$ and then de-average (lines 1 and 2). Next, calculate the covariance matrix *Cov_matrix*, find the eigenvalues of the matrix, and sort them (lines 3 and 4). Finally, the product of the matrix formed by the eigenvectors corresponding to the first N eigenvalues and the original matrix D is calculated to obtain the reduced-dimensional matrix O (lines 5 and 6).

The steps to implement compression encryption are as follows. Firstly, the one-dimensional word vectors are sorted by magnitude after PCA compression. Next, compare the privacy entities filtered by named entity recognition and get the index position of the corresponding word vector where the entity is located. Then the word represented by the word vector corresponding to the next index of the entity index position is returned and used to replace the private entity to achieve the encryption effect. The specific compression encryption effect is as follows in Fig. 3:

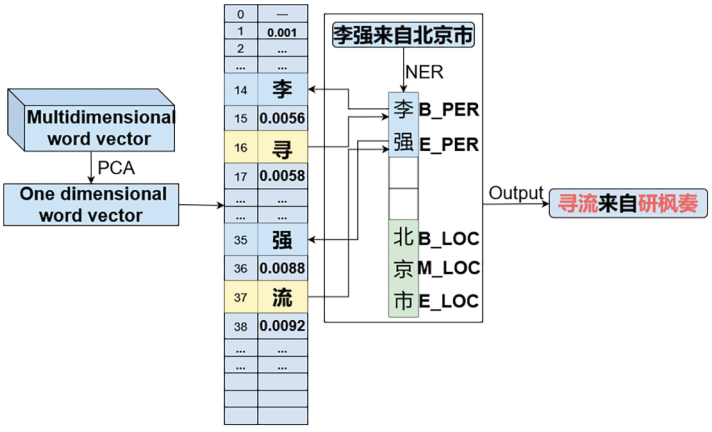

Figure 3: Compression and encryption effect

5 Experimental Comparison and Analysis

5.1 Experimental Data Processing

For the original dataset used in this paper, after converting the lexical annotation to BIEO format, the updated data became what is shown in Fig. 4:

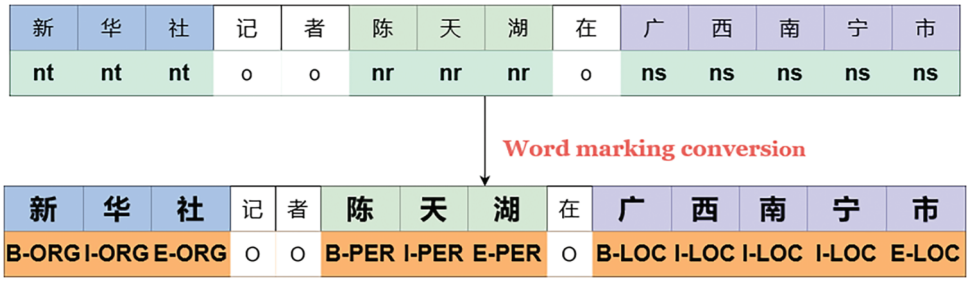

Figure 4: BIEO data labeling

Finally, after completing the annotation process of the People's Daily dataset, the dataset was divided into a training set and a test set according to the ratio of 9:1, and the statistics of the number of specific entities in each dataset are listed in the following Table 1:

Table 1: Number of entities statistics

	PER	LOC	ORG
Training set	17982	20193	9750
Test set	1999	2244	1084
Total	19981	22437	10834

5.2 The Setting of Important Parameters

In this experimental setting, the parameters of the model are gradually optimized to the best state after continuous training, and the training parameters of the model based on BiLSTM-CRF in this experiment are shown in Table 2. The experimental running platform is Pycharm.

Table 2: Experimental parameter setting

Parameter	Value
Hidden_dim	150
Batch_size	64
Embedding size	256
LR	0.001
Epoch	30

To evaluate the performance of the algorithm, evaluation metrics are proposed, Precision, Recall, and F1_score (i.e., a combined evaluation of Precision and Recall). True positive (TP) is the number of cases that would have been positive and were predicted to be positive, false positive (FP) is the number of

cases that would have been negative but were predicted to be positive, and false negative (FN) is the number of cases that would have been negative and were predicted to be negative, with the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F1\_score} = \frac{2 * \text{Precison} * \text{Recall}}{\text{Precison} + \text{Recall}} \tag{16}$$

5.3 Model Comparison Experiment

In this paper, not only the proposed BiLSTM-CRF model is experimented, we also compared the performance of the LSTM-CRF model, the BiLSTM model, and the traditional HMM model. The experimental results are compared in Table 3:

Table 3: Comparison of experimental results

Model	Precision	Recall	F1_score
LSTM-CRF	91.56%	90.99%	91.27%
BiLSTM	90.89%	73.43%	81.23%
HMM	77.86%	77.61%	77.74%
BiLSTM-CRF	**92.72%**	**92.30%**	**92.51%**

After 30 rounds of training, the deep learning-based BiLSTM-CRF is compared with the F1_score of the LSTM-CRF model and BiLSTM model in Fig. 5:

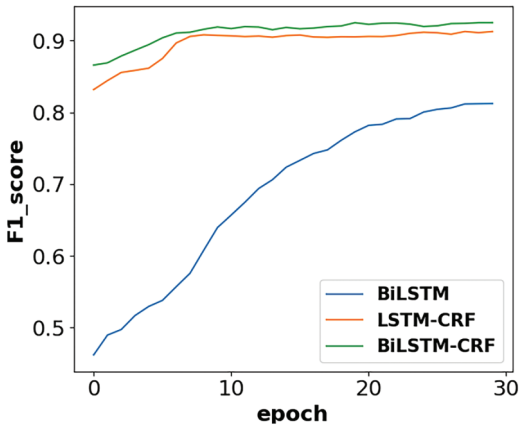

Figure 5: Model performance comparison

The performance comparison chart shows that both deep learning models are optimized and their performance improves as the number of training sessions gradually increases, and after 30 rounds of training, both models gradually converge. In the comparison graph of the performance of the neural network-based model, it can be visually observed that the performance of the model is improved after the

addition of CRF. The performance of the Bi-directional LSTM network gains further improvement compared to the one-way LSTM network. The HMM model is trained by generating a state transfer probability matrix, an observation probability matrix, and an initialization matrix, and then predicted with the help of the Viterbi algorithm.

Next, the loss values of the neural networks were also compared. For the BiLSTM model without the addition of CRF, the training loss values of the model after 30 rounds of training are shown in Fig. 6:

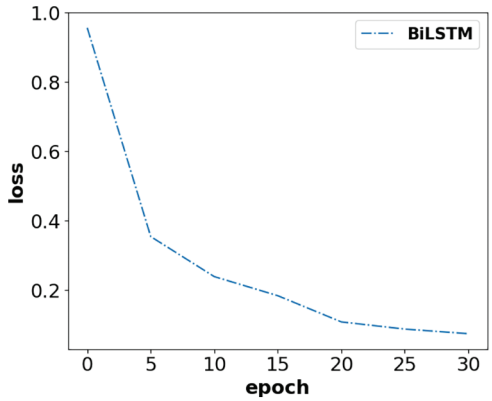

Figure 6: BiLSTM-loss

When CRF is added to the neural network model, the comparison of the BiLSTM-CRF model with the LSTM-CRF model for training loss is shown in Fig. 7:

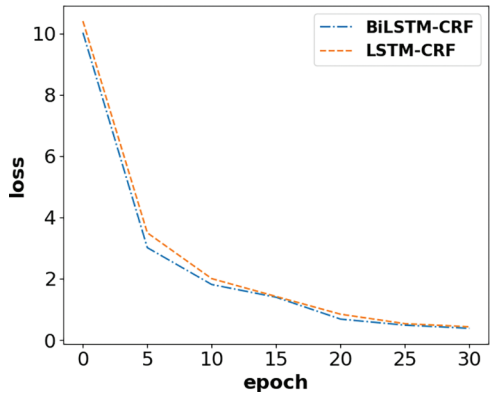

Figure 7: Comparison of BiLSTM-CRF model and LSTM-CRF model loss

For the model proposed in this paper, the process of training was also chosen to compare the performance of the model under different parameters by controlling variables. First, keep hidden_dim as 150 and adjust the size of LR. Through comparison experiments, we found that the model performs best when LR is set to 0.001. Second, a fixed LR is chosen so that the model performs best at 0.001, and then the size of hidden_dim is adjusted. It is found that the performance of the model improves with the increase of hidden_dim. However, too high is likely to cause overfitting, which reduces the performance of the model. So, the final choice of hidden_dim is 150. The comparative experiment is shown in Fig. 8 below.

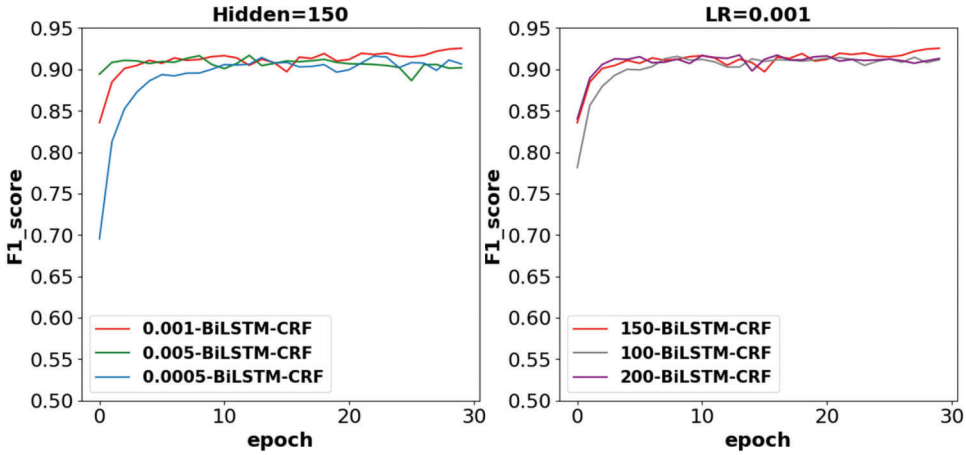

Figure 8: Comparison of different hidden_dim and LR

By comparing different batch_size sizes, the variation of model performance is obtained in Fig. 9:

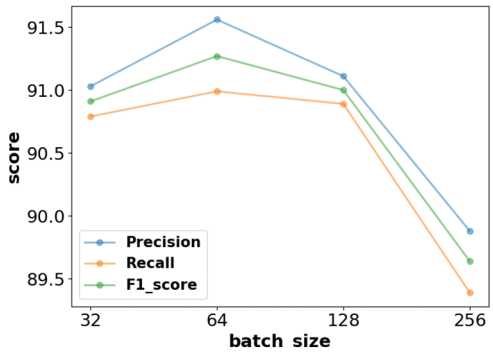

Figure 9: Comparison of different batch_size

Batch_size is the size of the data for each training session, and its size has an impact on the training time and performance of the model. Too large or too small batch_size will be detrimental to the training of the model. From the above Fig. 9, the precision, recall, and F1_score of the model show a rising and then falling trend as the batch_size increases. To obtain the best performance of the model, the final batch_size chosen for this model is 64.

To verify the superiority of the model used in this paper compared to other models, the experiments in Table 4 below were conducted. The experiments compared the prediction performance of the different models for different entity types.

By analyzing the data in Table 4, it has been found that compared to the LSTM-CRF model, the proposed model in this paper improves 1.0% (LOC), 1.5% (ORG), and 1.7% (PER) in F1_score, respectively. Compared to the BiLSTM model, the model used in the paper improved by 6.9% (LOC), 12.1% (ORG), and 13.7% (PER). Compared to the HMM model, the model used in the paper improved by 12.8% (LOC), 21.9% (ORG), and 13.0% (PER). To more visually compare the performance of the four models in named entity recognition, the F1_score values of the corresponding entities of the four models are converted into the following histogram:

Table 4: Controlled experiments with different entities

Model name	Entity category	Precision (%)	Recall (%)	F1_score (%)
	LOC	92.0	90.9	91.4
LSTM-CRF	ORG	87.8	89.1	88.4
	PER	92.8	92.6	92.7
	LOC	92.2	79.8	85.6
BiLSTM	ORG	90.5	68.1	77.8
	PER	91.0	72.1	80.5
	LOC	81.1	78.2	79.6
HMM	ORG	69.3	66.9	68.0
	PER	78.9	83.6	81.2
	LOC	**93.1**	**92.0**	**92.4**
BiLSTM-CRF	**ORG**	**89.2**	**90.6**	**89.9**
	PER	**94.6**	**93.8**	**94.2**

Through Fig. 10, it is obvious that the named entity recognition performance of the BiLSTM-CRF model based on Chinese electronic medical records has a considerable advantage in the recognition of different entity classes. This further confirms the effectiveness of the model for the recognition of named entities in the medical field. Deep learning combined with CRF will be more beneficial for NER tasks with contextual labels and sequence associations.

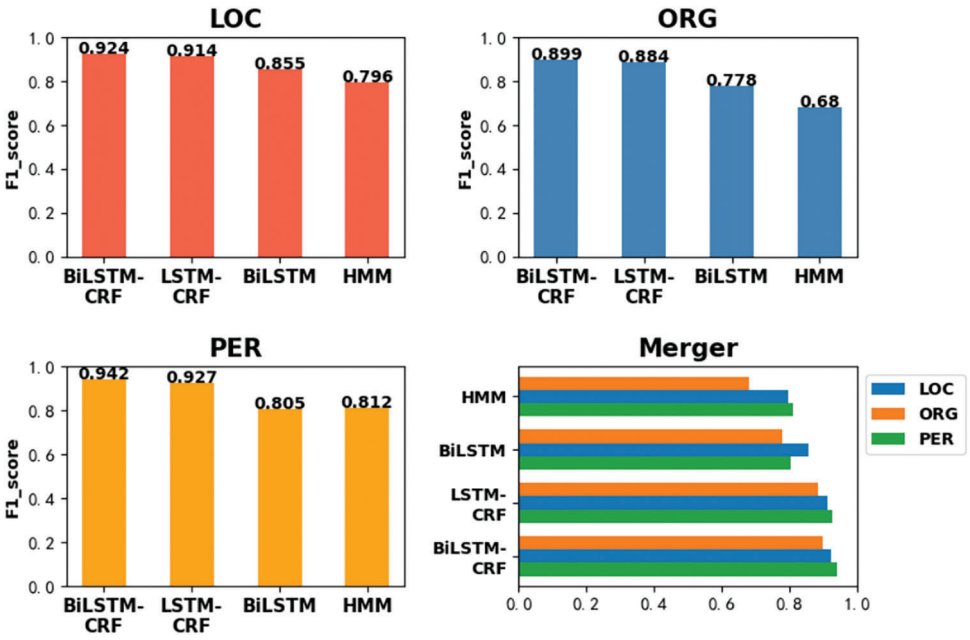

Figure 10: Comparison of F1_score of different entities

6 Conclusion

The BiLSTM-CRF model for named entity recognition and data masking of Chinese electronic medical records is proposed in this paper. It is to provide a feasible data masking and encryption method to complement the easy leakage of private data in the medical field. This experimental model combines regular expressions, and PCA word vector compression technology to bring more guarantees for data masking. This model was also compared with the LSTM-CRF-based model, BiLSTM-based model, and the HMM-based model with strict control of the parameter variables. And the performance differences between the different models were compared by comparing the precision values, recall values, and F1_score values. Through the final results, we found that the BiLSTM-CRF model is suitable for entity recognition tasks with context and label associations. It has great advantages in named entity recognition of electronic medical records privacy data, which plays a critical role in the privacy protection of medical data.

However, this study needs further optimization due to machine performance, time overhead, and other factors. Future work will continue to optimize the model, train a larger sample set to improve the applicability and compatibility of the model, and incorporate more new technologies to improve efficiency. Finally, the feasibility and reliability of using this model in other fields is also a future research direction and task.

Funding Statement: This research was supported by the National Natural Science Foundation of China under Grant (No. 42050102) and the Postgraduate Education Reform Project of Jiangsu Province under Grant (No. SJCX22_0343). Also, this research was supported by Dou Wanchun Expert Workstation of Yunnan Province (No. 202205AF150013).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Kong, G. Li, W. Rafique, S. Shen, Q. He *et al.,* "Time-aware missing healthcare data prediction based on arima model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022. https://doi.org/10.1109/TCBB.2022.3205064

[2] A. Agarwal, S. Sharma, V. Kumar and M. Kaur, "Effect of e-learning on public health and environment during COVID-19 lockdown," *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 104–115, 2021.

[3] R. K. Srivastava and D. Pandey, "Speech recognition using HMM and soft computing," *Materials Today: Proceedings*, vol. 51, pp. 1878–1883, 2022.

[4] B. Yang, W. Wu, Y. Liu and H. Liu, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[5] X. Xu, H. Tian, X. Zhang, L. Qi, Q. He *et al.,* "DisCOV: Distributed COVID-19 detection on X-ray images with edge-cloud collaboration," *IEEE Transactions on Services Computing*, vol. 15, no. 3, pp. 1206–1219, 2022.

[6] S. Kanwal, K. Malik, K. Shahzad, F. Aslam and Z. Nawaz, "Urdu named entity recognition: Corpus generation and deep learning applications," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 1, pp. 1–13, 2019.

[7] Y. Liu, Z. Song, X. Xu, W. Rafique, X. Zhang *et al.,* "Bidirectional GRU networks-based next POI category prediction for healthcare." *International Journal of Intelligent Systems*, vol. 37, no. 7, pp. 4020–4040, 2022.

[8] L. Kong, L. Wang, W. Gong, C. Yan, Y. Duan *et al.,* "LSH-Aware multitype health data prediction with privacy preservation in edge environment," *World Wide Web*, vol. 25, no. 5, pp. 1793–1808, 2022.

[9] T. Bi, X. Chen, J. Li and S. Yang, "Research on industrial data desensitization algorithm based on fuzzy set," in *2020 IEEE Int. Conf. on Advances in Electrical Engineering and Computer Applications (AEECA)*, Dalian, China, pp. 1–5, 2020.

[10] L. Qi, Y. Liu, Y. Zhang, X. Xu, M. Bilal *et al.,* "Privacy-aware point-of-interest category recommendation in internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21398–21408, 2022.

[11] W. Xu, Y. Xu, G. Huo, Y. Yang and Y. Jin, "Optimized dual-mode security encryption chip design based on hash algorithm," in *2022 IEEE 11th Int. Conf. on Communication Systems and Network Technologies (CSNT)*, Indore, India, pp. 566–570, 2022.

[12] A. Rehman, A. Khan, M. A. Ali, M. U. Khan, S. U. Khan *et al.,* "Performance analysis of PCA, sparse PCA, kernel PCA and incremental PCA algorithms for heart failure prediction," in *2020 Int. Conf. on Electrical, Communication, and Computer Engineering (ICECCE) IEEE*, Istanbul, Turkey, pp. 1–5, 2020.

[13] J. Duan, B. Wang, Z. Tan, X. Wei and H. Wang, "Chinese spelling check via bidirectional LSTM-CRF," in *2019 IEEE 8th Joint Int. Information Technology and Artificial Intelligence Conf. (ITAIC) IEEE*, Chongqing, China, pp. 1333–1336, 2019.

[14] K. Jiao and X. Li, "Extraction method of judicial language entities based on regular expression," in *2021 6th Int. Conf. on Intelligent Computing and Signal Processing (ICSP)*, Xi'an, China, pp. 372–376, 2021.

[15] G. Cenikj, B. Vitanova and T. Eftimov, "Skills named-entity recognition for creating a skill inventory of today's workplace," in *2021 IEEE Int. Conf. on Big Data*, Orlando, FL, USA, pp. 4561–4565, 2021.

[16] E. T. Khaing, M. M. Thein and M. M. Lwin, "Stock trend extraction using rule-based and syntactic feature-based relationships between named entities," in *2019 Int. Conf. on Advanced Information Technologies (ICAIT) IEEE*, Yangon, Myanmar, pp. 78–83, 2019.

[17] J. Zhang, M. Cui and B. Wang, "SAR image change detection method based on neural-CRF structure," in *2021 IEEE Int. Geoscience and Remote Sensing Symposium IGARSS IEEE*, Brussels, Belgium, pp. 3797–3800, 2021.

[18] Q. Guo, S. Wang and F. Wan, "Research on named entity recognition for information extraction," in *2020 2nd Int. Conf. on Artificial Intelligence and Advanced Manufacture (AIAM)*, Manchester, United Kingdom, pp. 121–124, 2020.

[19] W. Ting, "An acoustic recognition model for English speech based on improved HMM algorithm," in *2019 11th Int. Conf. on Measuring Technology and Mechatronics Automation (ICMTMA) IEEE*, Qiqihar, China, pp. 729–732, 2019.

[20] Y. Li, Q. Ma and X. Wang, "Medical text entity recognition based on CRF and joint entity," in *2021 IEEE Asia-Pacific Conf. on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, pp. 155–161, 2021.

[21] F. Yi, B. Jiang, L. Wang and J. Wu, "Cybersecurity named entity recognition using multi-modal ensemble learning," *IEEE Access*, vol. 8, pp. 63214–63224, 2020.

[22] V. K. Gupta, A. Gupta, D. Kumar and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 116–123, 2021.

[23] J. Hu and X. Zheng, "Opinion extraction of government microblog comments via BiLSTM-CRF model," in *Proc. of the ACM/IEEE Joint Conf. on Digital Libraries in 2020*, Virtual Event, China, pp. 473–475, 2020.

[24] J. Wang, "Study on named entity recognition in Chinese literatures on hypertension treatment," in *2021 the 3rd Int. Conf. on Intelligent Medicine and Health*, Macau, China, pp. 79–83, 2021.

[25] Z. Hu, X. Hu, L. Qi, S. Xue and X. Xu, "An information extraction method for sedimentology literature with semantic rules," in *2021 IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, Int. Conf. on Pervasive Intelligence and Computing, Int. Conf. on Cloud and Big Data Computing, Int. Conf. on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, AB, Canada, pp. 475–481, 2021.

[26] K. Zhang, D. Yue and L. Zhuang, "Improving Chinese clinical named entity recognition based on BiLSTM-CRF by cross-domain transfer," in *Proc. of the 2020 4th High Performance Computing and Cluster Technologies Conf. & 2020 3rd Int. Conf. on Big Data and Artificial Intelligence*, Qingdao, China, pp. 251–256, 2020.

[27] S. K. Duppati and A. R. Babu, "Deep learning based named entity recognition using bi-directional long short-term memory," *Journal of Optoelectronics Laser*, vol. 41, no. 7, pp. 746–53, 2022.

[28] Z. Wang and H. Guan, "Research on named entity recognition of doctor-patient question answering community based on BiLSTM-CRF model," in *2020 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM) IEEE*, Seoul, Korea (South), pp. 1641–1644, 2020.

[29] P. D. Waggoner, R. Y. Shapiro, S. Frederick and M. Gong, "Uncovering the online social structure surrounding COVID-19," *Journal of Social Computing*, vol. 2, no. 2, pp. 157–165, 2021.

[30] S. Dakrory, B. A. Abdelatif, M. Kayed and A. A. Ali, "Extracting geographic addresses from social media using deep recurrent neural networks," in *2021 9th Int. Japan-Africa Conf. on Electronics, Communications, and Computations (JAC-ECC)*, Alexandria, Egypt, pp. 135–139, 2021.

[31] Q. Liu, X. Jia, W. Yang, F. Tu and L. Wu, "Research on entity relation extraction based on BiLSTM-CRF classical probability word problems," in *2021 13th Int. Conf. on Education Technology and Computers*, Wuhan, China, pp. 62–68, 2021.

[32] C. Y. Kuo and E. J. L. Lu, "A BiLSTM-CRF entity type tagger for question answering system," in *2021 IEEE Int. Conf. on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) IEEE*, Bandung, Indonesia, pp. 161–166, 2021.

[33] R. Catelli, V. Casola, G. De Pietro, H. Fujita and M. Esposito, "Combining contextualized word representation and sub-document level analysis through bi-LSTM + CRF architecture for clinical de-identification," *Knowledge-Based Systems*, vol. 213, pp. 106649, 2021.

[34] L. Yan and S. Li, "Grape diseases and pests named entity recognition based on BiLSTM-CRF," in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conf. (IMCEC)*, Chongqing, China, pp. 2121–2125, 2021.

[35] L. Qi, Y. Yang, X. Zhou, W. Rafique and J. Ma, "Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6503–6511, 2022.

[36] R. R. Suman, B. Mondal and T. Mandal, "A secure encryption scheme using a composite logistic sine map (CLSM) and SHA-256," *Multimedia Tools and Applications*, vol. 81, pp. 27089–27110, 2022.

[37] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.