scientific reports

Check for updates

OPEN

Design of agricultural question answering information extraction method based on improved BILSTM algorithm

Ruipeng Tang[1]✉, Jianbu Yang[2], Jianxun Tang[3], Narendra Kumar Aridas[1] & Mohamad Sofian Abu Talip[1]

With the rapid growth of the agricultural information and the need for data analysis, how to accurately extract useful information from massive data has become an urgent first step in agricultural data mining and application. In this study, an agricultural question-answering information extraction method based on the BE-BILSTM (Improved Bidirectional Long Short-Term Memory) algorithm is designed. Firstly, it uses Python's Scrapy crawler framework to obtain the information of soil types, crop diseases and pests, and agricultural trade information, and remove abnormal values. Secondly, the information extraction converts the semi-structured data by using entity extraction methods. Thirdly, the BERT (Bidirectional Encoder Representations from Transformers) algorithm is introduced to improve the performance of the BILSTM algorithm. After comparing with the BERT-CRF (Conditional Random Field) and BILSTM algorithm, the result shows that the BE-BILSTM algorithm has better information extraction performance than the other two algorithms. This study improves the accuracy of the agricultural information recommendation system from the perspective of information extraction. Compared with other work that is done from the perspective of recommendation algorithm optimization, it is more innovative; it helps to understand the semantics and contextual relationships in agricultural question and answer, which improves the accuracy of agricultural information recommendation systems. By gaining a deeper understanding of farmers' needs and interests, the system can better recommend relevant and practical information.

Keywords Information extraction, Question and answer system, Natural language processing, Knowledge graph, Agricultural information recommendation

Now many farmers face some problems in agricultural production, such as climate change, market uncertainty, resource management, crop diseases, pest control, soil health, and food security, which affect the sustainability of agricultural production. In order to solve these problems, many farmers will search for solutions through the Internet[1]. However, the current data sources of agricultural knowledge are becoming more and more abundant, and the data volume is expanding[2]. The question-answering system is an intelligent information retrieval that answers natural language questions through dialogue. It can analyze a large amount of agricultural-related data (such as meteorological data, soil data, crop growth data, etc.), use natural language processing (NLP) technology to understand and process the questions raised by farmers, and provide accurate and useful answers. Some question-answering systems even integrate IoT devices, which can collect and analyze various farm data in real time, combining IoT and Internet data[3]. In order to ensure that the answers provided by the agricultural question-answering system are accurate, information extraction is the most important part. It can accurately extract key information from a large amount of agricultural data and literature, integrate these heterogeneous data together, standardize data of different formats and structures, and form a unified knowledge base for easy system retrieval and use[4].

In order to improve the performance of information extraction, some scholars have made some achievements. Lin et al.[5] proposed a joint neural framework OneIE. It extracts the global optimal IE result from the

[1]Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia. [2]Faculty of Languages and Linguistics, University of Malaya, 50603 Kuala Lumpur, Malaysia. [3]Faculty of Electronics and Electrical Engineering, Zhaoqing University, No. 55, Zhaoqing City, Guangdong Province, China. ✉email: 22057874@siswa.um.edu.my

input sentence as a graph and combines global features to capture cross-subtask and cross-instance interactions, achieving information extraction with global features. Luan et al.[6] proposed a general framework for information extraction using dynamic span graphs by selecting the most credible entity spans and linking these nodes with confidence-weighted relation types and coreferences. Nie et al.[7] proposed a connected system consisting of three homogeneous neural semantic matching models. They can perform document retrieval, sentence selection, and claim verification to extract and verify factual information. Yang et al.[8] routed instances to appropriate annotators to predict annotation difficulty to improve task routing and model performance for biomedical information extraction and improve the accuracy of information extraction.

Sahu et al.[9] used various inter-sentence and intra-sentence dependency constructions to capture local and non-local dependency information and proposed a sentence-to-sentence relation extraction model based on document-level graph convolutional neural network. Jiang et al.[10] established a path from the target to the candidate opinions through a directed syntactic dependency graph, and proposed a relational graph convolutional network based on the attention mechanism to utilize the syntactic information on the dependency graph to achieve goal-oriented opinion word extraction. Guo et al.[11] constructed a large self-annotated corpus in the field of agricultural pests and diseases, and proposed a new CNER model based on joint multi-scale local context features and self-attention mechanism (JMCA-ADP) for identifying named entities in the field of agricultural pests and diseases in China. Adnan et al.[12] proposed an unstructured multidimensional big data information extraction method to extract useful information from unstructured or semi-structured data. Chen et al.[13] used a deep learning algorithm to extract cultivated land information from agricultural remote sensing images to solve problems such as ambiguity of remote sensing data, large data volume, and large intra-class differences.

Although the above studies have made good achievements in information extraction in some fields, most methods rely on a large amount of high-quality annotated data, while agricultural data lacks sufficient annotated data support. In addition, agricultural data comes from various sources and has complex formats, so these methods are not capable of integrating and processing heterogeneous data. Some methods rely on multi-scale local context features, and have limited processing capabilities for long texts and complex contexts in the agricultural field. Although Padilla et al.[14] combined data from different sources and data mining techniques to extract the best association rules, and used spatial prediction technology to make more precise sales forecasts in geographic space, it relies on multi-source data fusion and complex spatial prediction technology, and the workload of data acquisition and preparation is large. In addition, technologies such as multivariate cokriging and association rule mining require high computing resources and time, and there is also a risk of overfitting. So this study proposes an agricultural information extraction method based on the BE-BILSTM algorithm, which is oriented to the question-answering scene of agricultural information retrieval. It can obtain the key information from massive question-answering data accurately, which is beneficial for the agricultural personnel to carry out in-depth data mining and application.

Methods and materials

The information extraction method proposed in this study has three modules: information crawling, information extraction rules and information extraction algorithm. Information crawling describes the process of information extraction. Information Extraction Rules describes the extraction rules of structured and unstructured data. The Information extraction algorithm introduces the Bert algorithm to improve the BILSTM algorithm and proposes the BE-BILSTM algorithm.

Information collection and extraction

Information crawling

This study extracts three types of information: soil type, crop diseases and pests, and agricultural trade, which is obtained from various official agricultural information websites in Malaysia. It has agricultural information networks, agricultural statistics databases, etc. This study also uses Python's Scrapy crawler framework[15], starts from the homepage of target websites, obtains URL[16] containing these three types of agricultural information, and removes invalid error messages, colloquial vocabulary, punctuation characters and other cleaning rules by removing stop words. Finally, it is converted into JSON (JavaScript Object Notation)[17] data and stored in the database. Figure 1 shows the scrapy crawling data process.

Data definition of information extraction rules

(1) Semi-structured information extraction

The information extraction of semi-structured data converts the website search result interface and its HTML (Hypertext Markup Language)[18] code into the structured information that can be stored in a relational database, which is < table > < /table > in the form. So the desired information can be extracted by judging the specific location of the web page element that is indicated in the XML (Extensible Markup Language)[19] web page path. Taking the web interface of crop diseases and pests as an example to realize the conversion between semi-structured data and relational database[20], it is necessary to set the extraction rules in the quadruple format as the actual situation of the disease and pest data. According to ontology and entity concepts of the information extraction, each entity belongs to an ontology layer concept. So each entity contains multiple attributes, each with an attribute value. Therefore, the establishment of collection $X = \{x_1, x_2, x_3, \ldots, x_j\}$ represents the ontology layer concept of the entity resource. For the source data set, $Y = \{y_1, y_2, y_3, \ldots, y_j\}$ represents the data set from entity resources to entity attributes, and

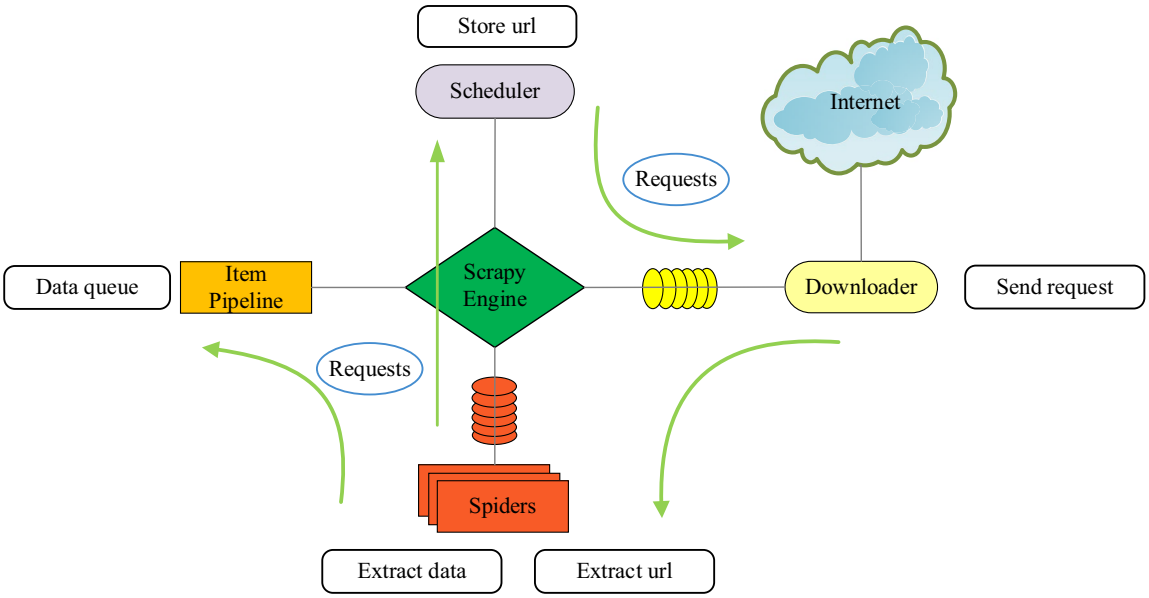

Fig. 1. The scrapy crawling data process.

$Z = \{z_1, z_2, z_3, \ldots, z_j\}$ represents the data set from entity attributes to attribute values, then the formula of crop diseases and pests data is shown in Formula 1.

$$Q = \{X, Y, Z, W\} \tag{1}$$

In Formula 1, Q represents the entity set of data, X represents the specific category of data, Y represents the entity information of data, Z represents the entity attribute of data, and W represents the value of the entity attribute of data. Information extraction rules for semi-structured data are applied to crop diseases and pests data, and the results are: < Corn insect pest, European corn borer, {scientific name, host, hazard characteristics, morphological characteristics, living habits, control methods, distribution area}, etc>. Figure 2 shows the information extraction process for the semi-structured data.

(2) Unstructured information extraction

Most unstructured data exists in text, and the information extraction for text data requires the completion of entity extraction[21]. In order to extract entities from text, the locations of entities have to be determined first, and other entities are determined to be placed in predefined categories by using the HMM (Hidden Markov Models)[22] method for the entity extraction. It focuses on transforming entity extraction problems into feature labeling problems for processing. Notes are not only related to current labels but also to predicted labels. Starting points, endpoints, and external entities of each word are obtained first through an inside-out labeling system. A model training dataset is built to match the attribute and classification of each word. Finally, a trained HMM model is taken to predict the entities of unstructured data. The structure of HMM model is shown in Fig. 3.

In Fig. 3, a_s represents the hidden state of the label at time s, b_s represents the hidden state of the label at time s. Formula 2 show that the annotation in the hidden state is only related to the state at the previous moment. Formula 3 show that the predicted value only relates to the current HMM chain. The successful operation of the model needs to satisfy the conditions of Formula 2 and 3, which are as follows:

$$G(a_s|a_{s-1}, a_{s-1}, \ldots, a_1, b_{s-1}, b_{s-1}, \ldots, b_1) = G(a_s/a_{s-1}) \tag{2}$$

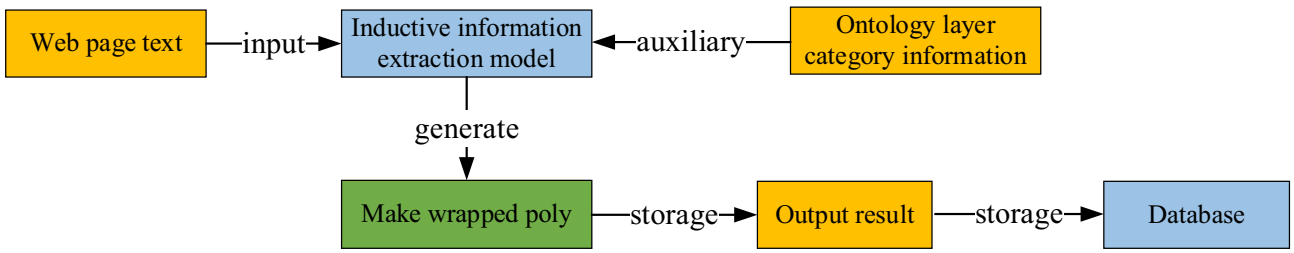

Fig. 2. The information extraction process for semi-structured data.

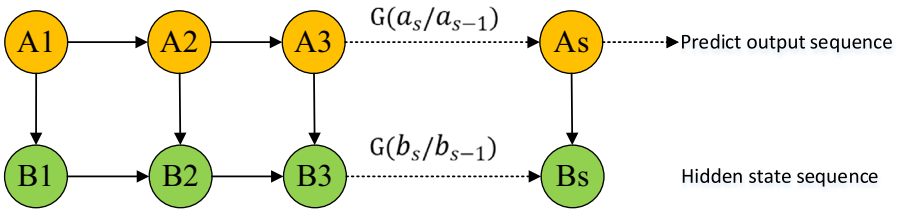

Fig. 3. The structure of HMM model.

$$G(b_s|a_{s-1}, a_{s-1}, \ldots, a_1, b_{s-1}, b_{s-1}, \ldots, b_1) = G(b_s/a_s) \tag{3}$$

Impact on RE-BILSTM model
The process of crawling information from web pages and extracting data from semi-structured and unstructured sources is crucial for the BE-BILSTM model's effectiveness in entity recognition for the following reasons:

(1) Diversity of data sources and model generalization

- *Diverse data sources* Web crawling provides access to a vast array of agricultural data sources, including agricultural news, research reports, and market analyses. These sources offer rich contextual information that helps the model better understand and recognize different agricultural entities.
- *Enhanced generalization* The diversity of data sources helps the model learn from a variety of contexts and scenarios, improving its generalization ability and robustness in real-world applications.

(2) Processing semi-structured and unstructured data

- *Semi-structured data* Information from HTML tables and structured web content provides clear entity and attribute information. For instance, parsing HTML tables can extract crop varieties, pest names, and control methods.
- *Unstructured data* Text data from web pages can capture implicit entity relationships and contextual information through natural language processing techniques. The BE-BILSTM model leverages the combination of BERT and BILSTM to achieve superior semantic representation and entity recognition performance when processing this unstructured data.

(3) Enhanced accuracy in entity recognition

- *Rich contextual information* The extracted information from diverse data sources provides rich contextual semantics, aiding the model in accurately recognizing entities. For example, combining textual descriptions with structured table data enhances the model's ability to identify crops and related pest information accurately.
- *Improved precision* By extracting and integrating information from different data formats, the BE-BILSTM model can better understand relationships between entities, leading to higher accuracy and precision in recognition.

Information extraction algorithm

BERT algorithm
In the study of semantic representation in language models, BERT (Bidirectional Encoder Representations from Transformers)[23] models released by Google use a unique network structure to make the evaluation tasks reach new index heights. BERT is similar to other deep learning models. It also performs self-supervised learning based on a large amount of text corpus and gets the semantic feature representation of each word. It provides simple and complex models ($Bert_{simple}$ and $Bert_{complex}$), and the corresponding parameters are as Formula 4,5:

$$Bert_{simple} : T = 12, U = 768, V = 12, P = 110M \tag{4}$$

$$Bert_{complex} : T = 24,\ U = 1024,\ V = 16,\ P = 340M \tag{5}$$

In Formula 4 and 5, T represents the number of layers of model networks, U represents the number of self-attention in multi-head attention, V represents the dimension of hidden layers, and P represents the total parameters.

The BERT model structure uses a bidirectional Transformer encoder structure[24], and the Transformer uses self-attention to build the text model, where D, E, and F represent word vector matrices, and l_e represents the vector dimension of data models. The relationship between each word in the sentence and the word at the current position is calculated when performing semantic representation. The Transformer network model can learn a

richer semantic expression than vectors such as word2vec[25]. In addition, Transformer also uses the Multihead mechanism and softmax level to increase the representation space of self-attention and improves the model's ability to focus on different positions, as shown in Formula 6, 7, 8:

$$\text{Attention(D, E, F)} = \text{softmax}\left(\frac{D \times E^K}{\sqrt{l_e}}\right) \times \text{F} \tag{6}$$

$$\text{Multihead(D, E, F)} = \text{Concat}\left(h_1, h_2, \ldots, h_j\right) \times U^{\rho} \tag{7}$$

$$h_\sigma = \text{Attention}\left(\text{D} \times U_\sigma^D, \text{E} \times U_\sigma^D, \text{F} \times U_\sigma^F\right) \tag{8}$$

Sequential features can not be extracted when using the self-attention mechanism. So Transformer uses the position embedding when processing temporal features, as shown in Formula 9,10:

$$\text{LOED(loc}, 2\varphi) = \sin\left(\text{loc}/10000^{2\varphi/model_h}\right) \tag{9}$$

$$\text{LOED(loc}, 3\varphi) = \cos\left(\text{loc}/10000^{2\varphi/model_h}\right) \tag{10}$$

In Formula 9 and 10, LOED(loc, 2φ) and LOED(loc, 3φ) represent two different positional encodings in the Transformer model(due to the inability of the self-attention mechanism to handle the sequence's order information, positional encodings are utilized to represent the position information of each element in the input sequence). loc represents the position within the input sequence, φ represents the dimension of the positional encoding, $model_h$ represents the hidden layer dimension of the model. sin and cos are used to generate these positional encodings. For the degradation problem in deep learning, the Transformer coding unit uses the layer normalization and residual network to solve it, as shown in Formula 11, 12, which achieves the purpose of using the context information of words to obtain better semantic representation.

$$\text{Layer}_{norm}\left(\varepsilon_\varphi\right) = \beta \times \frac{\varepsilon_\varphi - \rho_i}{\sqrt{\alpha_\varphi^2 + \tau}} + \gamma \tag{11}$$

$$\text{Feed}_{data} = \max(0, V_1 + \epsilon_1) \times V_2 + \epsilon_2 \tag{12}$$

In Formula 11, $\text{Layer}_{norm}\left(\varepsilon_\varphi\right)$ represents the parameters of layer normalization, which is a normalization technique used in neural networks that helps stabilize the training process and accelerate convergence. ε_φ represents the input to the network. β and γ represent the learning parameters that scale and translate normalized values. α_φ^2 and τ represent the mean and variance. In Formula 12, Feed_{data} represents the FNN (Feedforward Neural Network) parameters. In Transformer, FFN is a feedforward neural network used to process the features of each position.V_1, V_2 *and* ϵ_1, ϵ_2 represent two sets of linear transformation parameters.max(0, $V_1 + \epsilon_1$) represents the activation function ReLU (Rectified Linear Unit) used for the nonlinear transformation.

BILSTM algorithm

BILSTM is a type of RNN (Recurrent Neural Network)[26] that addresses the challenge of gradient vanishing and exploding during training. In this study, LSTM is to classify words, CRF (Conditional Random Field)[27], and continuously obtain restrictive rules from the training data to ensure the predicted labels. For BILSTM, neurons exhibit excellent memory functions, which can memorize historical sequences during learning. However, the unidirectional LSTM model can only learn the historical information but cannot learn information about future periods, so the bidirectional BILSTM appears. The BILSTM[28] adopts two neurons to obtain the semantic information from the forward and backward directions. It finally inputs the results of the two hidden layers into the same output layer by splicing. Figure 4 shows the BILSTM neural network structure.

In Fig. 4, $\underline{U}$, U represents the output of the hidden layer unit on the forward LSTM and backward LSTM at time t. R($\vec{U}$, U) represents the vectors concatenated by the hidden layer unit two directions. The output n_i of BILSTM is shown in Formula 13 (W represents the weight matrix, and δ_n represents bias):

$$n_i = \text{R}\left(\vec{U}, \overleftarrow{U}\right) + \delta_n \tag{13}$$

BE-BILSTM algorithm

In order to improve the performance of the BILSTM algorithm, the study uses the BERT algorithm to perform the semantic representation of words and input the word vector of the representation into BILSTM. After the forward and backward extraction of BILSTM, the feature output of BILSTM is combined with BERT vector and inputted to Softmax to judge the word label[29]. The feature at this time combines the feature representation of BERT and BILSTM algorithm, which is connected to the CRF layer to obtain the final label sequence. The formulas 14 and 15 are shown as followings:

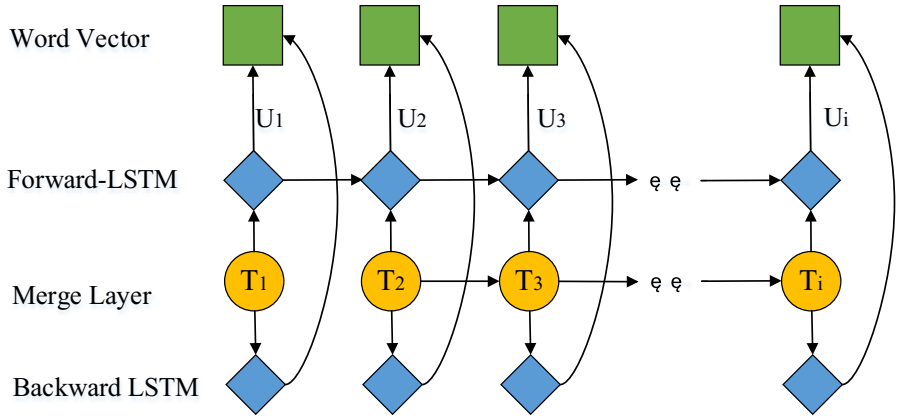

Fig. 4. The BILSTM neural network structure.

$$P(n|m) = \frac{\exp(score(m,n))}{O(m)} \tag{14}$$

$$O(m) = \sum n * \exp(score(m,n)) \tag{15}$$

In Formula 14 and 15, P(n|m) represents the probability of label n came from sentence m,O(m) represents the sum of indices scored by all label sequences *n*, the loss function is to calculate O(m), and the sequence with the highest transition probability for the min loss is the final prediction. Figure 5 shows the operation mechanism of BILSTM algorithm.

In Figure 5, the first step is the BERT word embedding. The input information is processed by the BERT algorithm; it uses its bidirectional encoder to generate contextual word embeddings for each word. The output of BERT is a two-dimensional matrix, with each word having a corresponding high-dimensional vector representation, which contains rich semantic information. The second step is the BILSTM feature extraction. The generated BERT word embedding is fed as input to the BILSTM layer. It extracts contextual features in the sequence through forward and backward LSTM units. The output of the BILSTM is a new feature matrix, with a feature vector containing forward and backward information for each time step. The third step is the feature concatenation. It concatenates the BERT word embedding vector with the BILSTM output vector to form a richer feature representation. The concatenated vector contains the contextual semantic information provided by BERT and the sequence dependency information captured by BILSTM. The four step is the CRF layer (conditional random field). The combined feature vector is input to the CRF layer. It generates the final label sequence by learning the label dependency in the sequence and performing global optimization. The output of the CRF layer is the label prediction for each word, taking into account the label consistency of the entire sequence. Through the above steps, the RE-BILSTM algorithm can fully utilize the advantages of BERT and BILSTM to improve the accuracy and efficiency of information extraction.

In this study, the main task of the BE-BILSTM algorithm is to extract useful information from agricultural question-answering data and convert it into a semi-structured format for subsequent retrieval and application. The following are the steps of the BE-BILSTM model in the specific information extraction and conversion process:

(1) Entity extraction.

- The question-answering data is input and the context-related word are generated through the BERT model.
- These word are input into the BILSTM for forward and backward sequence processing.

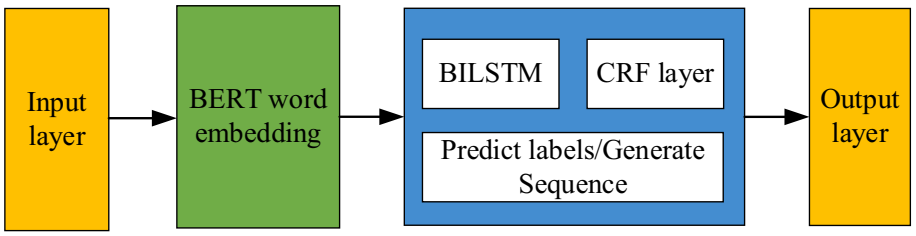

Fig. 5. The operation mechanism of BILSTM algorithm.

- The feature vector processed by BILSTM is used for label prediction through the CRF layer to extract entity information, such as crop type, pest name, agricultural product trading information, etc.

(2) Relationship extraction.

The feature vector extracted by the BILSTM model is used to further identify the relationship between entities. For example, identify the relationship between crops and their corresponding pests and diseases, or the price and quantity relationship in agricultural product transactions.

(3) Conversion to semi-structured information:

- The extracted entity and relationship information is formatted through predefined rules and templates. For example, convert crop and pest information into the JSON format, including fields such as crop name, pest name, hazard characteristics, and prevention and control methods.
- The transaction information is also converted into the JSON format, including fields such as agricultural product name, transaction price, quantity, etc.
- These semi-structured information can be conveniently stored in the database for subsequent query and analysis.

Experimental design

Data collection

This study collects 20,000 pieces of question-and-answer information about soil types, crop diseases and pests, and agricultural product trade from various agricultural websites in Malaysia(such as Malaysian agricultural information website, University Putra Malaysia agricultural information database, etc.) and the semantic database is constructed, which is divided into training and test sets according to the ratio of 7:3. Table 1 shows the content distribution of data set. Because the keywords corresponding to these three data types, so their attribute classification and identification keywords are defined. Table 2 shows some keywords corresponding to three types of information. In the above agricultural data, the BE-BILSTM model is mainly used to extract entities and relationships from agricultural question-answering data. These entities include crop types, pest names, agricultural product transaction information, etc. Relationships include the association between crops and pests, prices and quantities in transaction information, etc. The extracted entity and relationship information is formatted through predefined rules and templates and converted into semi-structured data in JSON (JavaScript Object Notation) format. JSON format is easy to store and retrieve, which can be easily integrated into the database or knowledge base.

Data annotation

In order to ensure the high accuracy of the RE-BILSTM model in agricultural question-answering information extraction, this study uses a pre-trained NER model to automatically annotate the initial data and manually correct the automatic annotation results. It uses the BIOES entity annotation method, where B represents the beginning, I represents the inside, O represents the outside, E represents that the corpus word is in the cutoff state, and S represents that the word can form a new entity. It also annotates agricultural data into three categories: soil, pests and diseases, and crops, forming the annotation set {soil, disease, crop}. Taking "Wet Clay Loam Soil", "Crucifer Downy Mildew", and "Shanghai Green" in the database as examples for annotation, the annotation results are shown in Table 3.

Model settings

In order to improve the experimental accuracy, this study sets the batch_size value (the number of samples for each training) to 32, and uses the search method to adjust the value of the learning rate during training

Question answering content	Total number of sentences (bars)	Number of training sets (bars)	Number of test sets (bars)
Soil type	8000	5600	2400
Crop pests	5000	3500	1500
Agricultural trade	7000	4900	2100
Total (bars)	20,000	14,000	6000

Table 1. The content distribution of the data set.

word attributes	Keyword example
Soil type	Soil, nitrogen, phosphorus, potassium, content…
Crop diseases and pests	Insects, diseases, habits, environment…
Agricultural trade	Price, kilogram, moisture, area…

Table 2. Part keywords correspond to three types of information.

Entity	Annotation information
Wet clay loam soil	Wet(B-soil), Clay(I-soil), Loam(O-soil), Soil(E-soil)
Crucifer downy mildew	Crucifer(B-disease), Downy(I-disease), Mildew (E-disease)
Shanghai green	Shanghai(B-crop), Green(E-crop)

Table 3. Part entity annotation examples.

continuously. It also uses the Adam optimizer to set the value to 0.0001, uses dropout to prevent overfitting and sets the value to 0.5. The number of iterations base_eproch is set to 100. The loss value of the model tends to a stable value of 0.08 as the number of iterations increases. Table 4 shows the experimental environment parameters.

Experimental results

In order to test the performance of BE-BILSTM algorithm, this study compares precision, recall and F1 values with BERT-CRF and BILSTM algorithm. The BERT-CRF algorithm[30] is used to extract context-related word vectors, and BERT is used to annotate the extracted words according to the dependencies of the labels. The experimental results are as follows:

(1) Overall model performance

Table 5 shows the overall model performance of three algorithms. In terms of Precision, Recall, and F1, the BE-BILSTM algorithm improves 2.96%, 6.23%, and 4.74% compared to the BERT-CRF algorithm. It also improves 4.81%, 7.00%, and 6.03% compared with the BILSTM algorithm. These result show that the performance of the BE-BILSTM algorithm is better. Figure 6 shows the F1 value trend of three algorithms. The training and prediction data diverge as the number of epochs reaches an inflection point around 6 rounds. The F1 value of BILSTM algorithm is lower than another two algorithms' values. Although BERT-CRF and BE-BILSTM algorithm are good in the training set and verification set, F1 value is more stable, which indicates that the BE-BILSTM algorithm can more accurately extract the key information from agricultural question-answering sentences.

(2) Class-specific model performance

In Figure 7, in terms of the Precision value in three types of agricultural information, the BE-BILSTM algorithm improves 3.93%, 1.91% and 4.81% compared to the BERT-CRF algorithm; it also improves 4.51%, 4.60% and 7.91% compared with the BILSTM algorithm. In terms of the Recall value of the three types of the agricultural information, the BE-BILSTM algorithm improves 10.16%, 7.39% and 5.64% compared with the BERT-CRF algorithm; it also improves 12.18%, 6.94% and 7.39% compared with the BILSTM algorithm. In the F1 value of the three types of agricultural information, the BE-BILSTM algorithm improves 9.15%, 8.85% and 4.87% compared with the BERT-CRF algorithm; it also improves 8.95%, 6.95% and 8.12% compared with the BILSTM algorithm. These results show that the performance of the BE-BILSTM algorithm in specific information processing is better than the BERT-CRF and BILSTM algorithms.

(3) Ablation model performance

Figure 8 and Table 6 show the ablation experimental results of three algorithms. The BERT algorithm can automatically learn contextual features, generate dynamic character vector representations, and perform well in agricultural information extraction tasks. In order to verify the impact of the BERT algorithm on the BE-BILSTM algorithm, this study introduced two other word embedding algorithms, Word2Vec[25] and GloVe to construct Word2Vec + BILSTM, GloVe + BILSTM, and BERT + BILSTM algorithm for ablation experiments. The Word2Vec algorithm represents words as vectors of continuous values, thereby mapping the vocabulary in

Tools/experimental environment	Version parameter
Python	3.8.12
Pytorch	1.8.2
Sklearn	0.26.1
Numpy	1.21.1
GPU RTX	3060
System	Centos7.4.0

Table 4. The experimental environment parameters.

Algorithm	Precision (%)	Recall (%)	F1 (%)
BERT-CRF	92.23	83.54	87.68
BILSTM	90.38	82.77	86.39
BE-BILSTM	95.19	89.77	92.42

Table 5. The overall model performance of three algorithms.

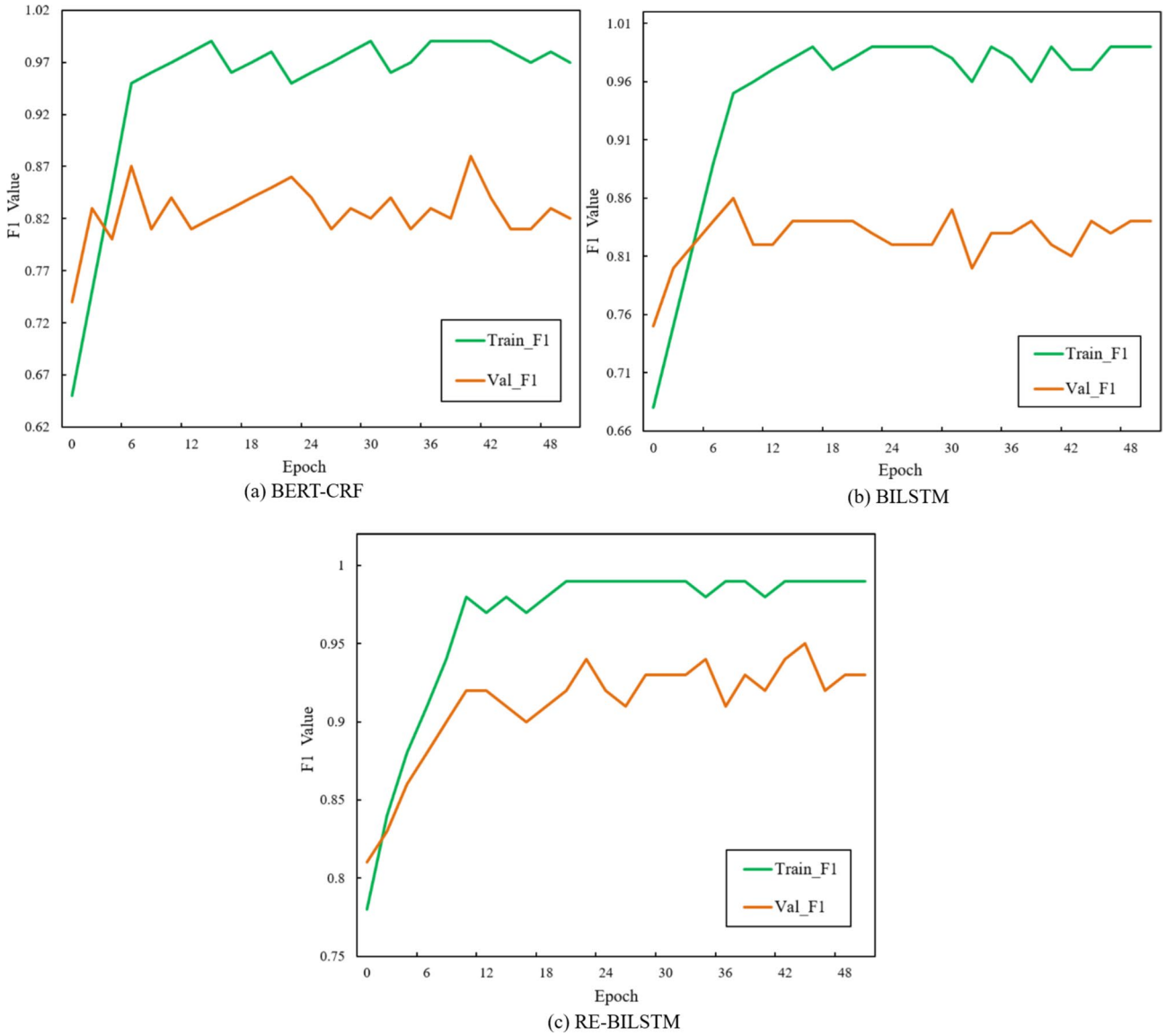

Fig. 6. The F1 value trend of different algorithm models.

the language into a low-dimensional continuous space so that computers can better understand and process the semantic relationships of words. The GloVe algorithm uses word co-occurrence statistics in the global corpus to learn the semantic relationships between words and generate embedding representations of words. The results are shown in Table 5. In terms of Precision, Recall, and F1, the BERT + BILSTM algorithm improves 6.53%, 4.20% and 5.35% compared to the Word2Vec + BILSTM algorithm. It improves 10.32%, 8.82% and 8.29% compared with the GloVe + BILSTM algorithm. These results show that the BERT algorithm performs better in agricultural information relationship extraction tasks, which can greatly improve the performance of the BILSTM algorithm.

Discussions

Although the BE-BILSTM algorithm proposed in this study can better extract agricultural question and answer information, some noise data is inevitably generated in the process of constructing the relationship extraction data set from word vectors. Although this study uses the target entity feature method to alleviates the impact of noisy words within sentences to a certain extent, but it ignores the impact of noise labels between sentences on the model. In subsequent study, other methods to alleviate noise samples such as reinforcement learning need to be used to improve the extraction performance of the model.

In addition, the current data collection method is too simple, and there is data imbalance in different data sets. There is a large gap in the sample size of different relationship types. Among the data used in the experiment, the sample size of the hazard relationship and part relationship is far more than that of other types. Therefore, agricultural information resources from different databases need to be further processed during the collection process by writing different data cleaning programs according to different data characteristics. However, this work is very tedious and requires detailed analysis of the type and structural characteristics of the data, which is

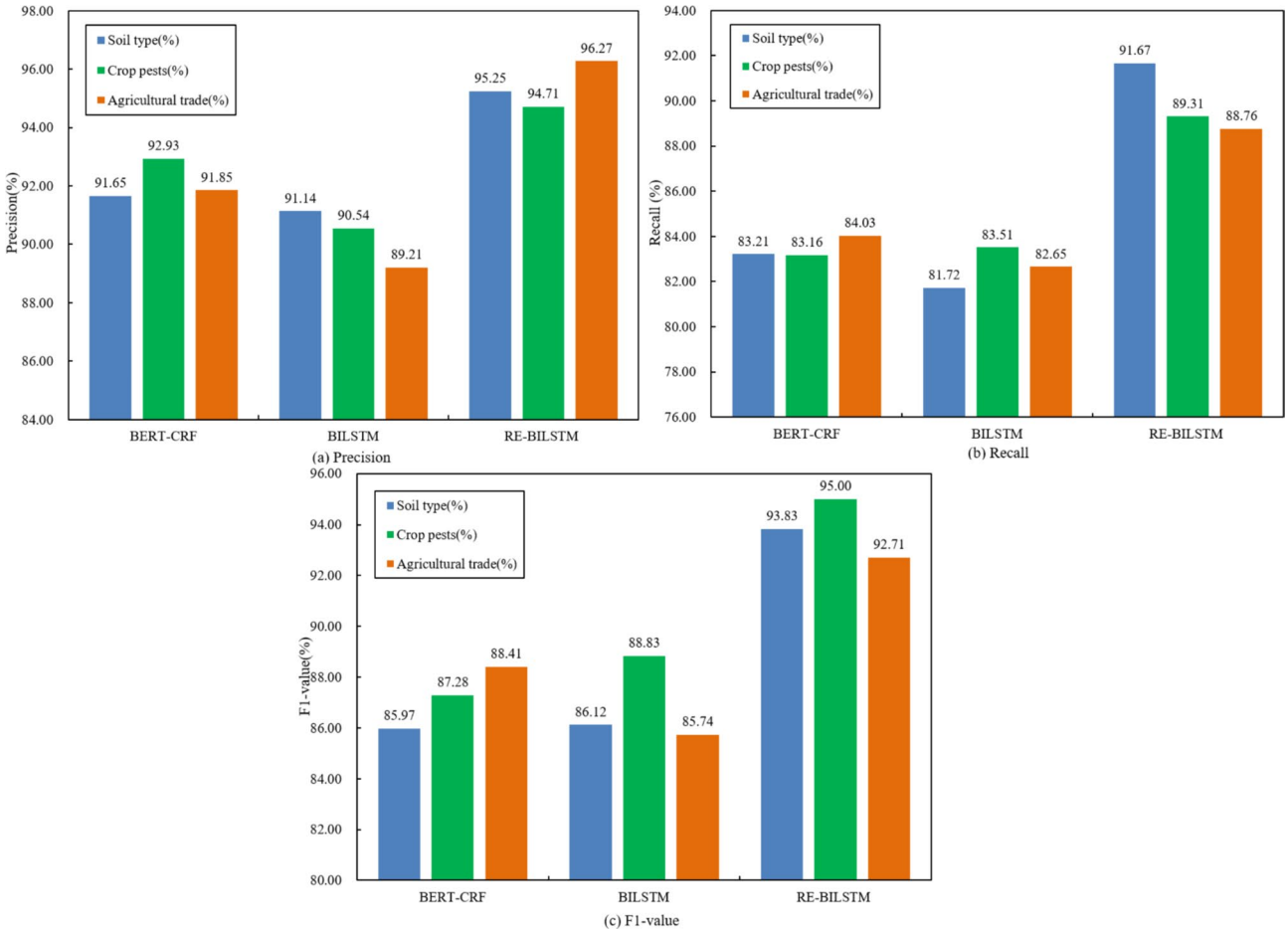

Fig. 7. The performance of three algorithms in extracting three types of agricultural information.

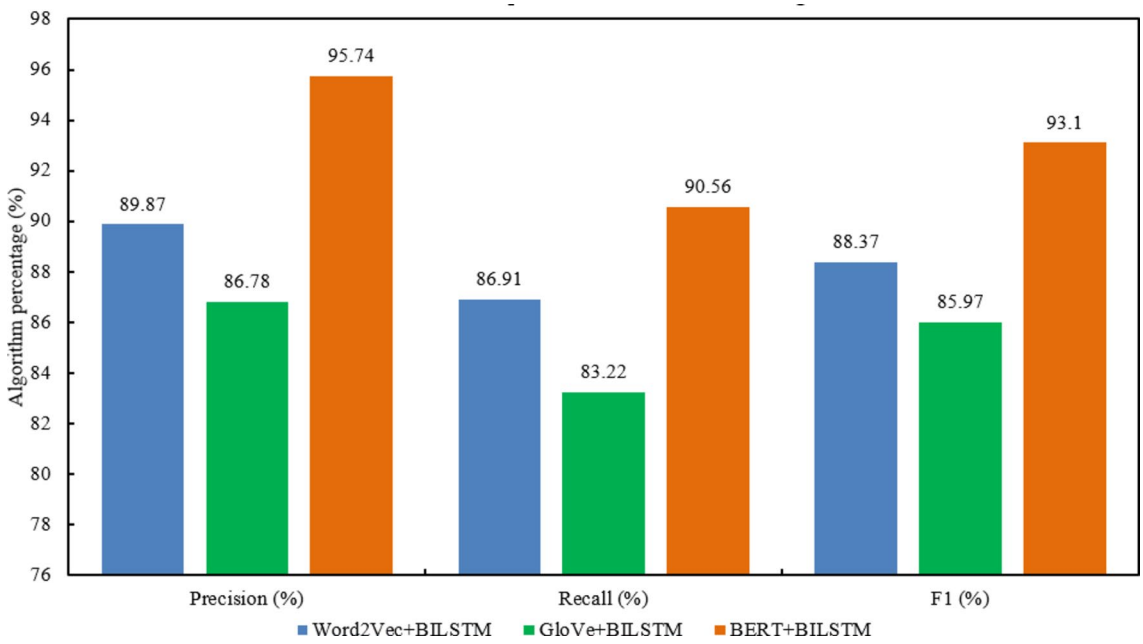

Fig. 8. The ablation experimental results of three algorithms.

Model	Precision (%)	Recall (%)	F1 (%)
Word2Vec + BILSTM	89.87	86.91	88.37
GloVe + BILSTM	86.78	83.22	85.97
BERT + BILSTM	95.74	90.56	93.10

Table 6. The ablation experimental results of three algorithms.

important for the future. Improving the recommendation accuracy of the agricultural resource recommendation system is a way that can be tried.

Finally, the BE-BILSTM algorithm proposed in this study only targets the currently used data sources for extracting unstructured data, which has great limitations. In the later expansion of the database, It will focus on learning different data based on artificial neural networks. The database characteristics can be used to extract knowledge from other data, thereby enhancing the scalability of the extraction method and better applying it to agricultural information recommendation systems. It can push useful information to users, which improves the efficiency and accuracy of users' acquisition in agricultural information to make better decisions.

Conclusions

In this study, an agricultural question-answering information extraction method based on the BE-BILSTM algorithm is designed. It uses Python's Scrapy tool, defines the extraction rules, introduces the BE-BILSTM algorithms to build the agricultural information extraction method. The experimental results show that the BE-BILSTM algorithm outperforms the BILSTM and BERT-CRF algorithms regarding accuracy, recall, and F1. The BERT algorithm has a better improvement effect on the BILSTM algorithm than the Word2Vec and GloVe algorithms. It shows that the BILSTM algorithm can more accurately identify and extract agricultural question-answering information, which provides users with efficient automatic question and answer services for agricultural knowledge. It can help accurately extract key information from agricultural question and answer information and reduce information distortion and misleading. This helps ensure the accuracy and quality of information, allowing farmers to obtain reliable agricultural knowledge and advice, improving agricultural production efficiency and sustainable development.

Data availability

All data generated or analyzed during this study are included in this published article and its Supplementary Information files, which are available from the corresponding author on reasonable request.

Received: 5 May 2024; Accepted: 19 August 2024
Published online: 18 October 2024

References

1. Li, C. & Niu, B. Design of smart agriculture based on big data and Internet of things. *Int. J. Distrib. Sens. Netw.* **16**(5), 1550147720917065 (2020).
2. Feng, X., Liu, Q., & Liu, X. Intelligent question answering system based on knowledge graph. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)* 1515–1520 (IEEE, 2021).
3. Valim Bandeira, M., Ferreira de Souza Móta, L. M. F., & Behr, A. Decision-making in agribusiness based on artificial intelligence. *Braz. J. Manag./Revista de Administração da UFSM* **15** (2022)
4. Yang, T., Mei, Y., Xu, L., Yu, H., & Chen, Y. Application of question answering systems for intelligent agriculture production and sustainable management: A review. *Resources, Conservation and Recycling* **204**, 107497 (2024).
5. Lin, Y., Ji, H., Huang, F., & Wu, L. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 7999–8009 (2020).
6. Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., & Hajishirzi, H. A general framework for information extraction using dynamic span graphs. Preprint at arXiv:1904.03296. (2019).
7. Nie, Y., Chen, H., & Bansal, M. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, No. 01, 6859–6866 (2019).
8. Yang, Y., Agarwal, O., Tar, C., Wallace, B. C., & Nenkova, A. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. Preprint at http://arxiv.org/abs/1905.07791 (2019).
9. Sahu, S. K., Christopoulou, F., Miwa, M., & Ananiadou, S. Inter-sentence relation extraction with document-level graph convolutional neural network. http://arxiv.org/abs/1906.04684 (2019).
10. Jiang, J., Wang, A., & Aizawa, A. Attention-based relational graph convolutional network for target-oriented opinion words extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* 1986–1997 (2021).
11. Guo, X. *et al.* Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Comput. Electron. Agric.* **179**, 105830 (2020).
12. Adnan, K. & Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. *J. Big Data* **6**(1), 1–38 (2019).
13. Chen, G., Sui, X., & Kamruzzaman, M. Agricultural remote sensing image cultivated land extraction technology based on deep learning. *Technology* **9**(10) (2019).
14. Padilla, W. R., García, J. & Molina, J. M. Knowledge extraction and improved data fusion for sales prediction in local agricultural markets. *Sensors* **19**(2), 286 (2019).
15. Fan, Y. Design and implementation of distributed crawler system based on Scrapy. In *IOP Conference Series: Earth and Environmental Science* vol. 108, no. 4, p. 042086 (IOP Publishing, 2018).

16. Wei, W. *et al.* Accurate and fast URL phishing detector: a convolutional neural network approach. *Comput. Netw.* **178**, 107275 (2020).
17. Bourhis, P., Reutter, J. L. & Vrgoč, D. JSON: Data model and query languages. *Inf. Syst.* **89**, 101478 (2020).
18. Ariyadasa, S., Fernando, S. & Fernando, S. Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML. *IEEE Access* **10**, 82355–82375 (2022).
19. Brahmia, Z., Hamrouni, H. & Bouaziz, R. XML data manipulation in conventional and temporal XML databases: A survey. *Comput. Sci. Rev.* **36**, 100231 (2020).
20. Kumar, A., Dabas, V. & Hooda, P. Text classification algorithms for mining unstructured data: A SWOT analysis. *Int. J. Inf. Technol.* **12**(4), 1159–1169 (2020).
21. Titouan, V., Courty, N., Tavenard, R., & Flamary, R. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning* 6275–6284 (PMLR. 2019).
22. Kalnoor, G. & Gowrishankar, S. A model for intrusion detection system using hidden Markov and variational Bayesian model for IoT based wireless sensor network. *Int. J. Inf. Technol.* **14**(4), 2021–2033 (2022).
23. Zheng, C., Deng, N., Cui, R., & Lin, H. Terminology extraction of new energy vehicle patent texts based on BERT-BILSTM-CRF. In *International Conference on Emerging Internetworking, Data & Web Technologies* 190–202 (Springer, 2023).
24. Zhang, D. C., Li, Z., Zhang, Y. & Lin, W. H. Noun metaphor recognition based on transformer and BERT. *Data Anal. Knowl. Disc.* **4**, 9 (2020).
25. Di Gennaro, G., Buonanno, A., & Palmieri, F. A. Considerations about learning Word2Vec. *J. Supercomput.* 1–16 (2021).
26. Saraswat, M. & Srishti,. Leveraging genre classification with RNN for book recommendation. *Int. J. Inf. Technol.* **14**(7), 3751–3756 (2022).
27. Zhou, X., Li, Y. & Liang, W. CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**(3), 912–921 (2020).
28. El Bourakadi, D., Ramadan, H., Yahyaouy, A. & Boumhidi, J. A novel solar power prediction model based on stacked BiLSTM deep learning and improved extreme learning machine. *Int. J. Inf. Technol.* **15**(2), 587–594 (2023).
29. Singla, P., Duhan, M. & Saroha, S. An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network. *Earth Sci. Inform.* **15**(1), 291–306 (2022).
30. He, X., Feng, J., Sun, F., Yan, M., Qian, J., Dai, W., & Wang, H. A Biomedical trigger word identification method based on BERT and CRF. In *International Conference on Web Information Systems and Applications* 393–402 (Springer, 2022).

Author contributions

Ruipeng Tang Conceptualization, R.T; Methodology, R.T; Software, R.T; Validation, R.T; Formal analysis, R.T; Investigation, R.T; Data curation, R.T; Writing—original draft, R.T; Writing—review & editing, R.T; Visualization, R.T; Jianbu Yang Conceptualization, J.Y; Methodology, J.Y; Validation, J.Y; Formal analysis, J.Y; Writing—review & editing, J.Y; Tang Jianrui Software, J.T; Investigation, J.T; Data curation, J.T; Visualization, J.T; Writing—original draft, J.T; Narendra Kumar Aridas Resources, N.K.A; Project administration, N.K.A; Writing—review & editing, N.K.A; Mohamad Sofian Abu Talip Writing—review & editing, M.S.A.T; Supervision, M.S.A.T; M.S.A.T; Resources, M.S.A.T; All listed authors have agreed to the manuscript content and the changes in the list of authors. We understand that these changes may require review and approval by the editorial board. All the above authors agree to be responsible for the content and conclusions of the article.

Funding

This research was funded by the laboratory itself and did not receive funding from external agencies.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.