agronomy

MDPI

Article

Named Entity Recognition of Chinese Crop Diseases and Pests Based on RoBERTa-wwm with Adversarial Training

Jianqin Liang [1], Daichao Li [1,*], Yiting Lin [1], Sheng Wu [1] and Zongcai Huang [2]

[1] The Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350116, China
[2] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
* Correspondence: lidc@fzu.edu.cn

Abstract: This paper proposes a novel model for named entity recognition of Chinese crop diseases and pests. The model is intended to solve the problems of uneven entity distribution, incomplete recognition of complex terms, and unclear entity boundaries. First, a robustly optimized BERT pre-training approach-whole word masking (RoBERTa-wwm) model is used to extract diseases and pests' text semantics, acquiring dynamic word vectors to solve the problem of incomplete word recognition. Adversarial training is then introduced to address unclear boundaries of diseases and pest entities and to improve the generalization ability of models in an effective manner. The context features are obtained by the bi-directional gated recurrent unit (BiGRU) neural network. Finally, the optimal tag sequence is obtained by conditional random fields (CRF) decoding. A focal loss function is introduced to optimize conditional random fields (CRF) and thus solve the problem of unbalanced label classification in the sequence. The experimental results show that the model's precision, recall, and F1 values on the crop diseases and pests corpus reached 89.23%, 90.90%, and 90.04%, respectively, demonstrating effectiveness at improving the accuracy of named entity recognition for Chinese crop diseases and pests. The named entity recognition model proposed in this study can provide a high-quality technical basis for downstream tasks such as crop diseases and pests knowledge graphs and question-answering systems.

Keywords: crop diseases and pests; named entity recognition; deep learning; pre-training language model; adversarial training

Citation: Liang, J.; Li, D.; Lin, Y.; Wu, S.; Huang, Z. Named Entity Recognition of Chinese Crop Diseases and Pests Based on RoBERTa-wwm with Adversarial Training. *Agronomy* **2023**, *13*, 941. https://doi.org/10.3390/agronomy13030941

Academic Editor: Roberto Marani

Received: 9 February 2023
Revised: 9 March 2023
Accepted: 20 March 2023
Published: 22 March 2023

1. Introduction

Crop diseases and pests (CDP) are an increasingly important factor threatening food security and restricting agricultural production [1]. To prevent and control the occurrence of CDP, obtaining CDP control information quickly and accurately is essential [2]. However, with the rapid development of information technology, the scale of CDP text data is growing exponentially. How to extract CDP knowledge from massive heterogeneous data sources has thus become an urgent problem. Crop disease and pest named entity recognition (CDP-NER) aim to accurately and efficiently identify entities related to CDP from massive unstructured data. This helps people obtain accurate, timely, and valuable information on disease and insect pest control and is of great significance to these challenges [3]. At the same time, CDP-NER is also the research basis for constructing CDP knowledge graphs and question-answering systems and directly affects the quality of these tasks [4,5]. Therefore, it is of great research value as well as practical significance for agricultural informatization in the study of effective entity recognition models in the field of CDP.

Named entity recognition (NER) [6] has been widely used in professional fields such as biomedicine [7], environmental science [8,9], and finance [10], but there are relatively few in-depth studies on Chinese named entity recognition of crop diseases and pests (CDP-CNER). In the past, the recognition of named entities of crop diseases and pests was mostly

accomplished via methods based on rules and dictionaries and machine learning [11]. Methods based on dictionaries and rules need to design rules or define dictionaries in advance, rely too much on experts, and are difficult to adapt to the continuous change and expansion of agricultural data in the era of big data [12]. Rule-based methods are gradually being replaced by machine learning methods. Models commonly used for entity extraction in the agricultural field include hidden Markov models (HMM) [13], maximum entropy Markov models (MEMM) [14], conditional random fields (CRF) [15], etc. Malarkodi et al. [16] proposed a CRF-based method of NER for real agricultural data. Although statistical machine learning methods have had some success in the field of CDP-NER, researchers must spend much energy on feature engineering and data annotation and contend with problems such as high-dimensional sparse data and poor scalability [17].

The development of deep learning [18] and artificial neural networks [19] has brought breakthroughs in the field of CDP-CNER. The deep learning method effectively solves the problems of traditional-approach NER reliance on artificial dictionaries and insufficient feature extraction [20]. Commonly used models include long short-term memory (LSTM) [21], recurrent neural networks (RNN) [22], and convolutional neural networks (CNN) [23] and their improved models [24]. Among them, the end-to-end BiLSTM [25] and CRF model that does not rely on artificial features has become the mainstream CDP-NER model and achieved promising results. However, because of the complex internal structure of LSTM, it takes much time and resources to train with the BiLSTM and CRF backbone models, so some researchers have attempted to simplify and improve LSTM, resulting in the gated recurrent unit (GRU) application [26]. In addition, these studies use traditional word vector models to obtain static representations and lack the ability to differentiate the same word with different meanings in different contexts. The large-scale pre-training model based on the Transformer model can represent polysemous words and has performed better in tasks such as CDP-NER [27]. Chen et al. proposed a named entity recognition model based on ALBERT to obtain word vectors and then input them into BiGRU-CRF for NER [28]. However, although introducing the pre-training model in the field of CDP-CNER has improved recognition accuracy to a certain extent, it can only be masked in units of words during pre-training. Word-level semantic information cannot be obtained, and fully learning complex text features in the field of CDP is challenging.

Overall, the field of CDP-CNER still faces many challenges. Firstly, because Chinese CDP texts are complex, they contain many complex terms, including the names of control agents such as "thiabendazole methyl wettable powder". At the same time, some entities express different semantics in different scenarios and may have unclear boundaries, creating a need to incorporate context and fully learn the characteristics of complex texts in the field of CDP. However, CDP-CNER still lacks the ability to obtain complex text features in the field of CDP and cannot obtain semantic information for Chinese word-level features. There is also the problem of uneven distribution of entity label samples in CDP data. For example, common entity categories such as "prevention and control chemicals" contain thousands of samples, while entity categories such as "etiology" only contain approximately 200 samples. The uneven distribution of samples seriously impacts overall entity recognition performance.

In order to solve the above problems, we propose a model that integrates the RoBERTa-BiGRU-CRF model and adversarial training (RGC-ADV). In RGC-ADV, dynamic vector representation at the word level of the input Chinese text information is obtained through the RoBERTa-wwm [29] pre-training model to solve the problems of incomplete word recognition and polysemy and fully learn the text features. At the same time, adversarial training is introduced to solve the problem of fuzzy boundaries. Entities with unclear boundaries are similar to input disturbances and may cause the model to make wrong decisions. Adversarial training is a method of adding confrontational disturbances to the word vector layer to enhance the robustness of the model to input disturbances [30,31]. A bidirectional GRU network is used to obtain connections among remote context semantics,

while the CRF layer at the output end is used to address dependency between labels and thus obtain a globally optimal label sequence. A focal loss function is then introduced to solve the problem of sample imbalance. Finally, three evaluation metrics, precision, recall [32], and F1 score [33,34], are used to evaluate the model.

The structure of this article is as follows. Section 2 introduces the CDP data set and the named entity recognition methods used in this study. Section 3 describes the experimental parameters and provides an analysis of the experimental results. Section 4 provides a discussion. Section 5 summarizes the research presented in this article.

2. Materials and Methods

2.1. Introduction to Data Set

Due to the lack of large-scale public annotation data sets for CDP and considering that China has more than a thousand crop resources [35], it would take a lot of labor and time to produce a corpus of all crops. Therefore, this paper selects Fujian Province of China as its research area. Based on the data of the Third National Crop Germplasm Resources Survey compiled by the Fujian Academy of Agricultural Sciences [36] and the Statistical Yearbook of Fujian Province [37], ten main crops are selected to obtain pest and disease materials, including five food crops, namely rice, soybean, wheat, barley, and sweet potato, and five cash crops, namely tea, sugar cane, peanut, radish, and rape.

2.1.1. Data Sources

There are many sources of knowledge in the CDP field. In this study, data published by the Crop Science Research Institute of the Chinese Academy of Agricultural Sciences [38] is selected as the main data source, while Baidu Encyclopedia is used as the supplementary data source. Auxiliary data sources include the CDP control experience of agricultural experts in Fujian Province as well as professional books such as *Application Manual of Technical Specifications for Major Crop Disease and Pest Prediction* [39] and *Atlas of Excellent Crop Germplasm Resources in Fujian Province* [40]. We determined our data acquisition method in relation to these different data sources. For the web data, we obtained corresponding URLs for the CDP data by parsing the web page structure and using regular expressions and BeautifulSoup to batch-parse the web page content. Published hard-copy data was scanned to form a PDF file, and optical character recognition (OCR) technology was then used to convert the PDF into text data. For the expert experience data, field surveys and expert consultation were adopted to obtain and sort out the text data. Finally, the various data sources were fused and aligned to obtain an original corpus for CDP. This contained a large amount of redundant data, so the original data were preprocessed before data annotation. Preprocessing included the deletion of invalid values, removal of inactive words, and the addition of missing values, resulting in a standardized CDP corpus with a total of about 170,000 Chinese characters.

2.1.2. Data Annotation

Based on the guidance of agriculture experts, eight entity categories were marked, namely diseases and pests, other names, etiology, damaged part, distribution areas, disease date, damaged crops, and prevention and control drug. The suffixes DIS, NAME, ETIOLOGY, PART, AREA, DATE, CROP, and DRUG were used to distinguish these categories. This paper uses the BIO labeling method, where B and I represent the beginning and the interior of the name of the CDP entity, while O represents the non-CDP entity part. The crop pest entity data labeling is shown in Table 1. The labeled data is divided into training, verification, and test sets in the ratio of 7:2:1.

Table 1. Crop disease and pest data annotation table.

Entity Name	Beginning Part	Inner Part	Other
Diseases and pests	B-DIS	I-DIS	O
Other names	B-NAME	I-NAME	O
Etiology	B-ETIOLOGY	I-ETIOLOGY	O
Damaged part	B-PART	I-PART	O
Distribution areas	B-AREA	I-AREA	O
Disease date	B-DATE	B-DATE	O
Damaged crops	B-CROP	I-CROP	O
Prevention and control drug	B-DRUG	I-DRUG	O

2.2. Proposed Approach

This paper proposes RGC-ADV as a CDP-CNER model that combines RoBERTa-wwm and adversarial training. The model structure is shown in Figure 1. The whole network can be divided into seven layers: input layer, RoBERTa-wwm layer, adversarial training layer, BiGRU layer, full connection layer, CRF layer, and output layer. First, the input text is pre-trained at the RoBERTa-wwm layer to obtain semantic information at the word level, convert each word of CDP text into a feature vector, and then the perturbation term is added to the vector representation to generate a counter sample. This last step improves the robustness of the model to input disturbance. The original vector representation and the adversarial sample are then input into BiGRU for training, fully learning the relationship between contexts. Last, the CRF layer is used to obtain the final prediction result, and a focal loss function is introduced to improve the problem of unbalanced CDP sample labels.

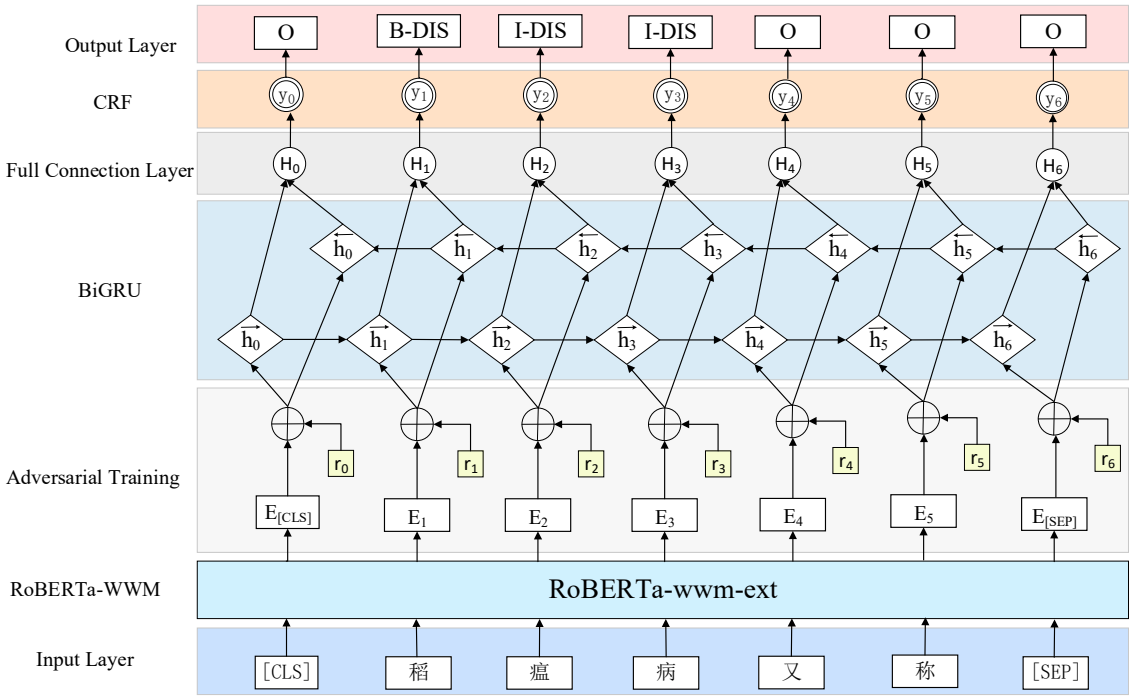

Figure 1. RGC-ADV named entity recognition model. Example " 稻瘟病又称" (this means another name for rice blast).

2.2.1. RoBERTa-wwm Model

RoBERTa-wwm (robustly optimized BERT pre-training approach-whole word masking) is based on BERT, a pre-training language model that uses a two-way transformer as an encoder to effectively fuse the information in the left and right parts of a word. BERT's

training goal is to accomplish next-sentence prediction (NSP) and generate a masked language model (MLM) [41]. The working principle of MLM is to randomly select 15% of words from the input disease and pest sentences for replacement, with an 80% probability of replacing them with a masking mark [MASK], 10% probability of replacing them with random words, and a 10% probability of keeping the original words unchanged. When executing the MLM task, BERT uses a static mask, with a random mask only applied once for each pest sample during the whole training process. RoBERTa, on the other hand, uses a dynamic mask, randomly selecting a certain proportion of words from the CDP sample for replacement in each iteration cycle. In this way, the model can obtain a greater representation of sentence patterns, thus improving the accuracy of the identification of named CDP entities for different diseases [42]. RoBERTa omits the NSP task and instead uses continuous input of Full-Sentences and Doc-Sentences until the maximum length of the input sentence is reached. Research has shown that RoBERTa improves the prediction of sentence relations. The wwm suffix [43] specifies that samples in the pre-training stage adopt the whole-word mask strategy and cover up with word granularity; this is effective for obtaining semantic representation at the word level. RoBERTa-wwm combines the advantages of RoBERTa and wwm. For the Chinese CDP corpus, BERT only covers a single Chinese character every time it executes MLM and cannot learn word-level semantic information. RoBERTa-wwm adopts the Chinese whole word masking. Firstly, the CDP corpus is segmented, and then the words are masked randomly. After covering all the Chinese characters that make up the same word, we predict these words. Finally, dynamic word vectors with word-level features are generated. It is more suitable for the task of CDP-CNER. The details are shown in Table 2.

Table 2. The masking strategy of BERT and RoBERTa-wwm. Example " 稻瘟病的症状" (this means the symptoms of rice blast).

Illustration	Sample
Original text	稻瘟病的症状
Segmented text	稻瘟病的症状
BERT's masking strategy	稻[MASK] 病的[MASK] 状
RoBERTa-wwm's masking strategy	[MASK] [MASK] [MASK] 的[MASK] [MASK]

This paper uses RoBERTa-wwm as the pre-training model for NER to extract text features. During the training process, the model parameters are fine-tuned according to the data provided by RoBERTa-wwm so that the model can better learn the semantic features of pest and disease data. RoBERTa-wwm is composed of 12-layer transformers, each using a multi-head attention mechanism to reduce the distance between two words at any position in the input pest sequence to a constant. The model structure is shown in Figure 2, where Tok_i, T_i, and E_i represent the ith word in the pest text data and the word vector before and after the transformer code. Suppose the input is $Tok = \left\{ Tok_{[cls]}, Tok_1, Tok_2 \ldots Tok_n, Tok_{[SEP]} \right\}$ while the vector $E = \left\{ E_{[cls]}, E_1, E_2 \ldots E_n, E_{[SEP]} \right\}$ corresponding to Tok, that is obtained through RoBERTa-wwm, contains the CDP semantic information obtained by RoBERTa-wwm in the pre-training stage.

2.2.2. Adversarial Training

Countermeasures training is used to perturb the model by adding some disturbance to the original input pest samples. This creates countermeasure samples which are then input into the model for training, effectively reducing noise due to personal information and improving the model's generalization ability. The research shows that the introduction of projected gradient descent (PGD) [44] can achieve an attack effect very close to the optimal global solution. Therefore, introducing PGD in the RGC-ADV model for iterative attack will consistently control the perturbation range within the specified range S. Once the dis-

turbance value exceeds the specified range S, it will fight against the sample x_t. Projecting to the specified range $x + S$, the iterative process is shown in Formula (1) [45]:

$$x_{t+1} = \prod_{x+S}(x_t + \alpha \cdot \mathrm{sign}(\nabla_x J(x_t, y))) \quad (1)$$

where α represents the size of disturbance in each iteration of PGD, $\prod_{x+s}$ is the projection operation, and x_t and x_{t+1} represent the adversarial samples generated during iteration steps t and t + 1. RGC-ADV obtains the initial vector E of CDP output from RoBERTa-wwm and adds perturbation to it to generate a CDP resistance sample E_{ADV}. The vectors E and E_{ADV} are trained together as the input of BiGRU.

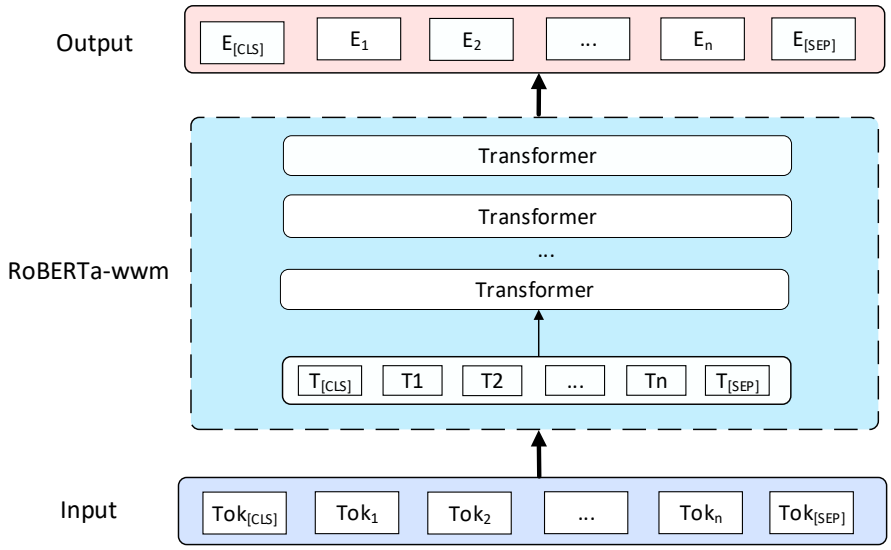

Figure 2. Structural diagram of RoBERTa-wwm model.

2.2.3. BiGRU Model

The GRU is developed based on LSTM and is composed of update and reset gates. The specific gate structure allows the GRU to solve the problem of gradient disappearance or explosion. Compared with LSTM, this has the advantages of fewer parameters, a simple structure, and low computational complexity. The internal GRU structure is shown in Figure 3. The update and reset gate are, respectively, composed of z_t and r_t; the update gate is used to determine the extent to which information from the previous time period is transmitted to the current period, while the reset gate is used to control how much information is forgotten. The main calculation process of the GRU network is shown in Formulas (2)–(5) [46]:

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \quad (2)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \quad (3)$$

$$\widetilde{h}_t = \tan h\left(W_{\widetilde{h}} \cdot [r_t * h_{t-1}, x_t]\right) \quad (4)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t \widetilde{h}_t \quad (5)$$

where x_t represents the pest information input at time t, σ indicates the sigmoid activation function, h_t and h_{t-1} represent the output vectors of the hidden layer at times t and $t-1$, $\widetilde{h}_t$ represents the state of the current candidate set, W_r, W_z, and $W_{\widetilde{h}}$ represent the input weight matrix of the activation function, $*$ represents the Hadamard product, and tan h is the activation function.

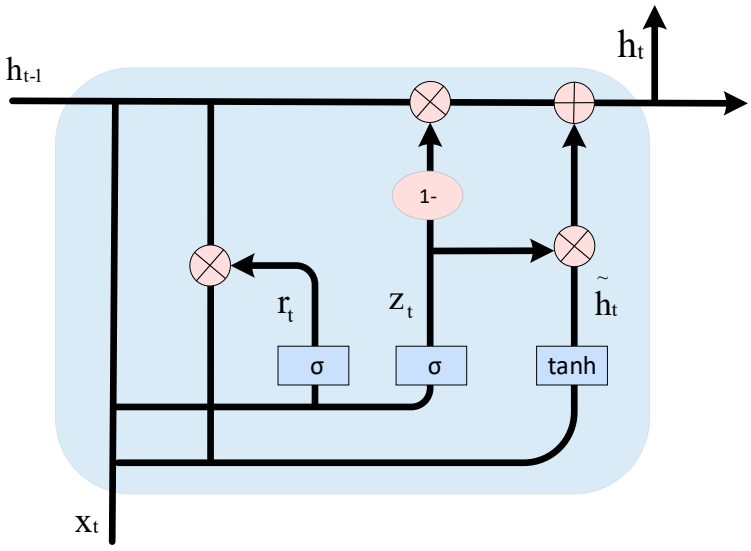

Figure 3. Internal unit structural diagram of GRU model.

However, because the GRU network can only process CDP data in one direction, it can only make predictions by acquiring the forward text data features [47]. Therefore, RGC-ADV uses BiGRU to train the CDP text output vector from the anti-training layer, capturing the semantic dependency of the context of CDP information and obtaining forward and backward data features. It uses these features to improve the accuracy of the predicted value [48].

2.2.4. Full Connection Layer

The full connection layer usually appears behind feature extraction operations such as the convolution layer and activation function and principally maps the learned distributed feature representation to the sample label space [49], thus classifying the samples. RGC-ADV first integrates the sample output feature results by BiGRU through the full connection layer, effectively weakening the impact of location features on the classification results. This improves the classification effect for the CDP samples.

2.2.5. Conditional Random Fields Model

CRF is a discriminant model for the correlation between labels that takes into account the transfer characteristics among CDP labels and uses CRF decoding to obtain the highest-probability group of label sequences. The score of the output sequence H is calculated by using the output sequence P from the full connection layer as the input to CRF. The calculation process is shown in Formula (6):

$$\mathrm{Score}(\mathrm{H}, \mathrm{y}) = \sum_{i=0}^{n} \mathrm{A}_{\mathrm{y_i y_{i+1}}} + \sum_{i=1}^{n} \mathrm{P}_{\mathrm{i, y_i}} \quad (6)$$

where n is the sequence length, A is the transfer fraction matrix, and $\mathrm{A_{ij}}$ represents the transfer score matrix element from the ith tag to the jth tag of the pest sample. When decoding, a Viterbi algorithm [50] is used to calculate the sequence tag of sequence H with the highest probability among all $\mathrm{H_{max}}$, as shown in Formula (7):

$$\mathrm{H_{max}} = \mathrm{argmax}(\mathrm{score}(\mathrm{H}, \mathrm{y})). \quad (7)$$

Adjacent CDP tags are ordered relationships in the task of CDP-CNER. For example, the I-DISEASE tag should appear after the B-DISEASE tag. However, the ability of BiGRU to deal with the dependency of learning tags is limited. Therefore, RGC-ADV adds a CRF layer to obtain the globally optimal tag prediction sequence. In addition, because of the unbalanced classification of the CDP sample labels, the focal loss function is introduced to

optimize the CRF model [51]. Focal loss makes the training process pay more attention to negative classification samples by controlling the weights of positive and negative samples to continuously optimize the performance of the model, as shown in Formula (8) [52]:

$$\mathrm{Loss}_{\mathrm{Focal}} = -\alpha(1 - \mathrm{P}(\mathrm{y}|\mathrm{x}))^{\gamma} \ln(\mathrm{P}(\mathrm{y}|\mathrm{x})) \quad (8)$$

where $\alpha \in [0, 1]$ is a factor used to balance the number of positive and negative samples in the pest samples; $\gamma \geq 0$ is the modulation coefficient, which is used to reduce the loss of easily classified (non–pest entity) samples, and make the model pay more attention to difficult (pest entity) samples; and $\mathrm{P}(\mathrm{y}|\mathrm{x})$ represents the probability that the label of the pest sample x is y.

3. Results

3.1. Experimental Parameter Setup

This study uses the RoBERTa-wwm model to pre-train the data set, ultimately obtaining a 768-dimensional vector representation. An AdamW optimizer is used to train the model, while the warmup learning rate strategy is used to assist learning. The initial learning rate is set to 0.001, the maximum length of the set sequence is 128, the hidden-layer dimension is 768, the batch size is set to 64, and the epoch is 30 rounds. In order to mitigate the impact of unbalanced label classification, the focal loss function is integrated into the CRF layer in order to optimize it with balance factor $\alpha = 0.96$ and $\gamma = 2$.

For model evaluation, this study uses the most commonly used evaluation indicators in the NLP field, namely Precision, Recall, and F1 score, to evaluate the model. Precision refers to the sample probability for all the samples predicted to be CDP. Recall refers to the sample probability that is accurately predicted from the sample of the actual pest entity. The F1 score is calculated by weighting the Precision and Recall rates and is a comprehensive reflection of the model evaluation results. The specific calculations are as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (9)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (10)$$

$$\mathrm{F1\ score} = \frac{2{*}\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \quad (11)$$

where TP means the correctly predicted sample number of CDP entities, TP + FP means the total number of samples of CDP entities predicted. TP + FN means the total sample number of CDP entities in the data set.

3.2. Experimental Results

3.2.1. Comparative Model Results

In order to verify the superiority of the RGC-ADV model proposed in this paper, a variety of models were used for comparative analysis. Two types of comparative experiments were performed: using different embedding methods and using different downstream model structures under the same embedding method. The selected models were BiGRU-CRF [53], BERT-BiGRU-CRF [54], ALBERT-BiGRU-CRF [55], RoBERTa-wwm-BiGRU-CRF, and RoBERTa-wwm-CRF [56]. The results are shown in Table 3.

Generally, the entity extraction effect of the RGC-ADV model proposed in this paper has certain advantages over other methods. The Precision rate is 89.23%, the Recall rate is 90.90%, and the F1 value is 90.04%. This shows that RGC-ADV is well adapted to the entity recognition task in the field of CDP and is effective at extracting the text data for this

field. From the results of the two groups of comparative experiments, we can draw the following conclusions:

1. Effectiveness of the embedding method:

Using the same data set and downstream model, the RoBERTa-wwm embedding method can identify CDP entities more accurately than ALBERT or BERT. RGC-ADV outscored BiGRU-CRF, BERT-BiGRU-CRF, and ALBERT-BiGRU-CRF in Precision, Recall, and F1 score by 9.15, 1.57, and 3.39 percentage points, 9.76, 0.33, and 5.36 percentage points, and 9.48, 0.97, and 4.4 percentage points, respectively. This shows that RoBERTa-wwm embedding can improve the ability of the model to perform semantic representation of the text, thus optimizing the effect of better CDP-CNER tasks;

2. Effectiveness of the downstream model:

Through the introduction of adversarial training, RGC-ADV's scores are increased by 0.67 percentage points in Precision, 1.24 percentage points in Recall, and 0.95 percentage points in F1 score. This improved performance shows that adversarial training can help the model better adapt within the CDP field. Moreover, compared with the RoBERTa-wwm-CRF model, RGC-ADV scores increased by 4.2, 4.19, and 4.19 percentage points in Precision, Recall, and F1, respectively. This further verifies that the proposed model can achieve strong entity recognition results in the field of CDP and has distinct advantages over other approaches.

Table 3. Comparison of experimental results for different entity recognition models.

Experiment Content	Model	Evaluating Indicator		
		Precision (%)	**Recall (%)**	**F1 Score (%)**
Other embedding methods	BiGRU-CRF	80.08	81.14	80.56
	BERT-BiGRU-CRF	87.66	90.57	89.07
	ALBERT-BiGRU-CRF	85.84	85.54	85.64
Our method	RoBERTa-wwm-adv-BiGRU-CRF (RGC-ADV)	89.23	90.90	90.04
Other downstream models	RoBERTa-wwm-BiGRU-CRF	88.56	89.66	89.09
	RoBERTa-wwm-CRF	85.03	86.71	85.85

3.2.2. Results for RGC-ADV Model

The RGC-ADV model proposed in this paper is used to train the CDP data set; model performance testing is carried out on eight types of entities. The experimental results are shown in Table 4. In general, RGC-ADV is effective at learning the feature information of CDP texts and has good recognition capability. The F1 scores for six types of entity recognition, namely diseases and pests, other names, etiology, damaged part, damaged crops, and prevention and control drug, are all more than 89%. This may be because descriptions are relatively simple for these entities, and the data characteristics are obvious. The recognition effect for the damaged part is weak, with an F1 score of 78.29%. We found that the description of the same crop part in the pest text is diverse, for instance, stem, stem base, stalk, and other words that all describe the plant stem. This makes it difficult to distinguish the entity boundary of the crop part. The recognition effect for the damage date is also poor. Here, the F1 score of 81.21% is related to the small number of release date samples and the relative difficulty of identifying the entity boundary; this leads to difficulty in the full learning of the model.

Table 4. Crop disease and pest entity recognition results.

Entity	Precision (%)	Recall (%)	F1 Score (%)
Diseases and pests	95.16	92.91	94.02
Other names	91.82	90.68	91.25
Etiology	98.51	98.85	98.68
Damaged part	76.23	80.46	78.29
Distribution areas	93.50	97.74	95.57
Disease date	79.76	82.72	81.21
Damaged crop	91.10	92.80	91.94
Prevention and control drug	87.76	91.02	89.36

4. Discussion

In recent years, improving CDP-CNER performance has been a research hotspot, and it is also a challenging and active area in the intelligent upgrading of CDP control. In this paper, the RGC-ADV model is applied to the CDP-CNER, aiming to solve the impact of the unbalanced distribution of entity samples and the indistinct entity boundary on the effect of CDP-CNER and solve the problem that Chinese word-level semantic information cannot be obtained in CDP-CNER. The model proposed in this paper has a good performance in the process of CDP-CNER.

The performance improvement of CDP-CNER has always been a problem that needs to be solved urgently [57]. At present, CDP-CNER is mainly based on pre-training language models to obtain semantic information. The word vector representations obtained by traditional word vector models such as Word2Vec [58] and GloVe [59] are static and cannot represent the ambiguity of words. Due to the complexity of CDP, it is difficult to use the traditional word vector models to obtain static representations to meet the actual needs of CDP-CNER. The pre-training language models based on Transformer is a dynamic text representation method, which will dynamically adjust the text representation according to the current crop pest and disease context [60]. Zhang et al. [61] completed the CDP-CNER based on the BERT pre-training model, effectively solving the problem of polysemy. However, when the BERT model pre-trains the Chinese CDP corpus, it can only cover up Chinese characters and not obtain word-level semantic information [62]. Liu et al. [63] proposed a NER model for wheat diseases and pests based on the fusion of ALBERT and rules. ALBERT reduces the number of model parameters compared to BERT and improves the ability and training speed of the model in acquiring sentence-level features. However, ALBERT still lacks the ability to acquire Chinese word-level features [64]. It is still unable to effectively solve the problem of incomplete word recognition and fuzzy boundaries in the field of CDP-CNER. The RoBERTa-wwm model not only inherits the advantages of BERT but also improves on it in terms of data volume and model structure. In addition, the Chinese full-word masking technology is used to realize the word-based masking method and obtain a vector representation with word-level semantic information [65]. In view of the above problems, this paper fully considers the text characteristics of CDP and uses the advantages of RoBERTa-wwm to solve the problem of incomplete recognition of CDP entity. Combined with adversarial training, it can effectively improve the accuracy of CDP-CNER and solve the problem of difficult recognition of entities with unclear boundaries to a certain extent.

The RGC-ADV model is effective in the field of CDP-CNER, which can help us better obtain CDP entities in text data. However, the model has a poor recognition effect on nested entities. The next stage of research will introduce more dimensional language features to enhance the recognition of nested entities of CDP. In addition, the coverage and scale of the CDP labeling corpus are small, which to a certain extent, restricts the improvement of entity recognition performance in the field of CDP.

5. Conclusions

This paper proposes a CDP-CNER model, RGC-ADV, which significantly improves the accuracy of CDP-CNER. The model effectively integrates RoBERTa-wwm-BiGRU-CRF and adversarial training for CDP-CNER to complete the optimal annotation of CDP text sequences. The main innovations are reflected in the following three aspects. First, the model fully considers the features of the corpus as well as the hidden features in the sentence. Through RoBERTa-wwm, a word vector representation that integrates the information and characteristics of CDP can be generated to alleviate the deviation caused by incomplete representation of semantic features in the prediction of the model and to enhance the semantic representation ability of the model for Chinese CDP text information. Second, adversarial training is introduced in the training process. Adversarial perturbation is added to the word vector layer to further improve the recognition effect of entities with unclear boundaries while improving the model's generalization ability. Third, a focal loss function is introduced to optimize the CRF; this effectively alleviates the impact of the unbalanced classification of the CDP label samples. Our results show that RGC-ADV exhibits a strong ability to recognize CDP entities in the CDP-CNER process, laying a solid foundation for CDP knowledge graphs and question-answering system tasks. It also provides a new research perspective for NER in garden plants, animals, and other fields. In the future, we will introduce richer feature information based on the RGC-ADV model to improve the recognition accuracy of CDP nested entities. At the same time, we will further explore the impact of different coverage and scale of CDP annotation corpus on CDP-CNER tasks.

Author Contributions: Conceptualization, D.L., J.L. and Y.L.; methodology, D.L. and J.L.; validation, D.L. and J.L.; formal analysis, Y.L. and J.L.; writing—original draft preparation, D.L. and J.L.; writing—review and editing, Z.H., D.L. and J.L.; supervision, Z.H., D.L. and S.W.; project administration, D.L. and S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Project, the Spatio-temporal Knowledge Graph and Knowledge Service for Crowd-Based Cooperative, grant number 2022YFB3904205.

Data Availability Statement: The data of this work can be shared with the readers depending on the request.

Acknowledgments: We thank the editors of Agronomy and the anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Athanassiou, C.G.; Phillips, T.W.; Wakil, W. Biology and Control of the Khapra Beetle, Trogoderma granarium, a Major Quarantine Threat to Global Food Security. *Annu. Rev. Entomol.* **2019**, *64*, 131–148. [CrossRef] [PubMed]
2. Zhao, J.S. Construction and Application of Knowledge Map of Crop Diseases and Pests Based on ALBERT. Master's Thesis, Anhui Agricultural University, Hefei, China, 2022.
3. Fountas, S.; Espejo-Garcia, B.; Kasimati, A.; Mylonas, N.; Darra, N. The Future of Digital Agriculture: Technologies and Opportunities. *IT Prof.* **2020**, *22*, 24–28. [CrossRef]
4. Zhao, S.; Luo, R.; Cai, Z.P. Overview of Chinese Named Entity Recognition. *Comput. Sci. Explor.* **2022**, *16*, 296–304.
5. Drury, B.; Roche, M. A survey of the applications of text mining for agriculture. *Comput. Electron. Agric.* **2019**, *163*, 104864. [CrossRef]
6. Grishman, R.; Sundheim, B. Message Understanding Conference—6: A Brief History. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5-9 August 1996; pp. 466–471.
7. Sung, M.; Jeong, M.; Choi, Y.; Choi, Y.; Kim, D.; Lee, J.; Kang, J. BERN2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **2022**, *38*, 4837–4839. [CrossRef]
8. Huang, Z.C.; Chou, P.Y.; Lucky, F.; Wu, S. Detection of environmental pollution events contained in online news text based on joint theme features. *J. Geoinf. Sci.* **2019**, *21*, 1510–1517.
9. Huang, Z.C.; Chou, P.Y.; Wang, H.B.; Wu, S. Typhoon event information extraction method combining event and context features. *J. Surv. Mapp. Sci. Technol.* **2019**, *36*, 209–214.

10. Xu, Q.R.; Zhu, P.; Luo, Y.F.; Dong, Q.W. Research progress of Chinese named entity recognition in financial field. *J. East China Norm. Univ. (Nat. Sci. Ed.)* **2021**, *5*, 1–13.
11. Wang, C.Y.; Wang, F. Research on agricultural named entity recognition based on conditional random field. *J. Hebei Agric. Univ.* **2014**, *37*, 132–135.
12. Kanwal, S.; Malik, K.; Shahzad, K.; Aslam, F. Urdu Named Entity Recognition: Corpus Generation and Deep Learning Applications. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *19*, 1–13. [CrossRef]
13. Lawrence, R.; Biinghwang, J. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16.
14. Mccallum, A.; Freitag, D.; Pereira, F. Maximum Entropy Markov Models for Information Extraction and Segmentation. *ICML* **2000**, *17*, 591–598.
15. Lafferty, J.; Mccallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. ICML* **2001**, *2001*, 282–289.
16. Malarkodi, C.; Lex, E.; Devi, S. Named Entity Recognition for the Agricultural Domain. *Res. Comput. Sci.* **2016**, *117*, 121–132.
17. Georgescu, T. Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents. *Symmetry* **2020**, *12*, 354. [CrossRef]
18. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
19. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall PTR: USA, 1998. Available online: https://dl.acm.org/doi/abs/10.5555/521706 (accessed on 13 March 2022.).
20. Zhao, R.X.; Yang, C.X.; Zheng, J.H.; Li, J.; Wang, J. Agricultural Intelligent Knowledge Service: Overview and Future Perspectives. *Smart Agric. (Chin. Engl.)* **2022**, *4*, 105–125.
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
22. Elman, J.L. Finding Structure in Time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
23. Lecun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation.* **1989**, *1*, 541–551. [CrossRef]
24. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
25. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Vancouver, BC, Canada, 26–31 May 2013.
26. Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwemk, H.; Bengio, Y. *Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation*; Association for Computational Linguistics (ACL): Doha, Qatar, 2014.
27. Wang, H.F.; Ding, J.; Hu, F.K.; Wang, X. Survey on large scale enterprise-level knowledge graph practices. *Comput. Eng.* **2020**, *46*, 1–13.
28. Chen, X.L.; Tang, L.Y.; Hu, Y.; Jiang, F.; Peng, L.; Feng, X.C. The extraction method of knowledge entities and relationships of landscape plants based on ALBERT model. *J. Glob. Inf. Sci.* **2021**, *23*, 1208–1220.
29. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
30. Miyato, T.; Dai, A.M.; Goodfellow, I. Adversarial Training Methods for Semi-Supervised Text Classification. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
31. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
32. Jones, S.; KAREN. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
33. Zhang, X.; Li, C.Z.; Du, H.C. Named Entity Recognition for Terahertz Domain Knowledge Graph based on Albert-BiLSTM-CRF. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 2602–2606.
34. Li, H. *Statistical Learning Methods 2nd Edition*; Tsinghua University Press: Beijing China, 2019.
35. Guo, Y.Y. *Institute of Plant Protection, Chinese Academy of Agricultural Sciences, China's Crop Diseases and Insect Pests*; China Agricultural Publishing House: BeiJing, China, 2015; p. 1746.
36. Lin, S.S.; Lu, P.L.; Zhang, H.F.; Ge, C.B.; Chen, S.L. Characteristic analysis of crop planting area and yield change in Fujian Province. *China Seed Ind.* **2022**, *8*, 73–79.
37. Statistical Yearbook of Fujian Province. Available online: https://tjj.fujian.gov.cn/xxgk/ndsj/ (accessed on 10 March 2022.).
38. China Crop Germplasm Information Network—Crop Disease and Pest Knowledge Website. Available online: https://www.cgris.net/disease/default.html (accessed on 13 March 2022).
39. National Agricultural Extension Service Center. *Application Manual of Technical Specifications for Major Crop Diseases and Pests Prediction*; China Agricultural Publishing House: Beijing, China, 2010.
40. Yu, W.Q. *Atlas of Excellent Crop Germplasm Resources in Fujian Province*; China Agricultural Publishing House: BeiJing, China, 2022.
41. Yang, P.; Dong, W.Y. Chinese Named Entity Recognition Method Based on BERT Embedding. *Comput. Eng.* **2020**, *46*, 40–45.
42. Zhang, Y.Q.; Wang, Y.; Li, B.C. Chinese electronic medical record named entity recognition based on RoBERTa-wwm dynamic fusion model. *IEEE* **2022**, *6*, 242–250.

43. Cui, Y.M.; Che, W.X.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
44. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
45. Li, Q.; Min, C.H.; Yang, Y.L.; Shen, C.; Fang, L.M. Deep learning model against attack and defense in full cloud edge scenario. *Comput. Res. Dev.* **2022**, *59*, 2109–2129.
46. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
47. Zhao, S.; Zhang, Y.; Wang, S.; Zhou, B.; Cheng, C. A recurrent neural network approach for remaining useful life prediction utilizing a novel trend features construction method. *Meas. J. Int. Meas. Confed.* **2019**, *146*, 279–288. [CrossRef]
48. Liu, S.; You, S.; Zeng, C.; Yin, H.; Liu, Y. Data source authentication of synchrophasor measurement devices based on 1D-CNN and GRU. *Electr. Power Syst. Res.* **2021**, *196*, 107207. [CrossRef]
49. Cheng, X.; Zhang, Y.; Chen, Y.; Wu, Y.; Yue, Y. Pest identification via deep residual learning in complex background. *Comput. Electron. Agric.* **2017**, *141*, 351–356. [CrossRef]
50. Li, W.J.; Zhang, Q.Q.; Zhang, P.Y.; Yan, Y.H.; Bai, L. Deep neural network speech endpoint detection based on Viterbi algorithm. *J. Chongqing Univ. Posts Telecommun. (Nat. Sci. Ed.)* **2018**, *30*, 210–215.
51. Guo, Y.B.; Li, Y.F.; Chen, Q.L.; Fang, C.; Hu, Y.Y. Network threat intelligence entity extraction integrated with Focal Loss. *J. Commun.* **2022**, *43*, 85–92.
52. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dolliar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
53. Jiao, Z.; Sun, S.; Ke, S. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv* **2018**, arXiv:1807.01882.
54. Qin, Q.; Zhao, S.; Liu, C. A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records. *Complexity* **2021**, *2021*, 1–11. [CrossRef]
55. Li, X.L.; Deng, Q.K. Chinese Position Segmentation Based on ALBERT- BiGRU-CRF Model. In Proceedings of the 2021 International Symposium on Computer Technology and Information Science (ISCTIS), Guilin, China, 4–6 June 2021; pp. 116–120.
56. Li, Z.; Cheng, N.; Song, W. Research on Chinese Event Extraction Method Based on RoBERTa-WWM-CRF. In Proceedings of the 2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 20–22 August 2021.
57. Nismi Mol, E.A.; Santosh Kumar, M.B. Review on knowledge extraction from text and scope in agriculture domain. *Artif. Intell. Rev.* **2022**, 1–43. [CrossRef]
58. Mikolov, T.; Chen, K.; Corrado, G.; Jeffrey, D. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
59. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
60. Li, Z.J.; Fan, Y.; Wu, X.J. A Survey of Research son Pretraining Technology for Natural Language Processing. *Comput. Sci.* **2020**, *47*, 162–173.
61. Zhang, W.H.; Wang, C.; Wu, H.; Zhao, C.J.; Teng, G.F.; Huang, S.F.; Liu, Z. Research on the Chinese Named-Entity–Relation-ExtractionMethod for Crop Diseases Based on BERT. *Agronomy* **2022**, *12*, 2130. [CrossRef]
62. Wang, J.Q.; Yu, L.; Xia, W.Y.; Feng, Q.; Wu, S.Z.; Chen, Z.P.; Fan, H.W.; Wu, Y. Named Entity Recognition Method in Power Network Dispatching Domain Based on ERNIE-IDCNN-CRF Model. *Power Inf. Commun. Technol.* **2022**, *20*, 8.
63. Liu, H.B.; Zhang, D.M.; Xiong, S.F.; Ma, X.M.; Xi, L. Named Entity Recognition of Wheat Diseases and Pests fusing ALBERT and Rules. *J. Front. Comput. Sci. Technol.* **2022**. [CrossRef]
64. Wen, C.D.; Zeng, C.; Ren, J.W.; Zhang, G. Patent text classification combined with ALBERT and two-way gated circulation unit. *Comput. Appl.* **2021**, *41*, 407–412.
65. Wang, S.Y.; Yuan, K. Emotional analysis model of college students' forum based on RoBERTa-WWM. *Comput. Eng.* **2022**, *48*, 292–298.