Hindawi
Journal of Sensors
Volume 2022, Article ID 3967171, 15 pages
https://doi.org/10.1155/2022/3967171

Research Article

TBR-NER: Research on COVID-19 Text Information Extraction Based on Joint Learning of Topic Recognition and Named Entity Recognition

Xin Feng,[1,2,3] **Yingrui Li**,[4] **Zhang Hang**,[4] **Zhang Fan**,[4] **Qiong Yu**,[1] **and Ruihao Xin**[4,5]

[1]*Department of Epidemiology and Biostatistics, School of Public Health, Jilin University, Changchun, Jilin 130012, China*
[2]*School of Science, Jilin Institute of Chemical Technology, Jilin 130000, China*
[3]*State Key Laboratory of Inorganic Synthesis and Preparative Chemistry, College of Chemistry, Jilin University, Changchun, Jilin 130012, China*
[4]*College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 130000, China*
[5]*College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China*

Correspondence should be addressed to Qiong Yu; yuqiong@jlu.edu.cn and Ruihao Xin; xrh@163.com

Received 28 December 2021; Revised 29 May 2022; Accepted 6 July 2022; Published 4 August 2022

Academic Editor: Ying-Ren Chien

There is a centralization of the core content in the text information of the new crown epidemic notification. This paper proposes a joint learning text information extraction method: TBR-NER (topic-based recognition named entity recognition) based on topic recognition and named entity recognition to predict the labeled risk areas and epidemic trajectory information in text information. Transfer learning and data augmentation are used to solve the problem of data scarcity caused by the initial local outbreak of the epidemic, and mutual understanding is achieved by topic self-labeling without introducing additional labeled data. Taking the epidemic cases in Hebei and Jilin provinces as examples, the reliability and effectiveness of the method are verified by five types of topic recognition and 15 types of entity information extraction. The experimental results show that, compared with the four existing NER methods, this method can achieve optimality faster through the mutual learning of each task at the early stage of training. The optimal accuracy in the independent test set can be improved by more than 20%, and the minimum loss value is significantly reduced. This also proves that the joint learning algorithm (TBR-NER) mentioned in this paper performs better in such tasks. The TBR-NER model has specific sociality and applicability and can help in epidemic prediction, prevention, and control.

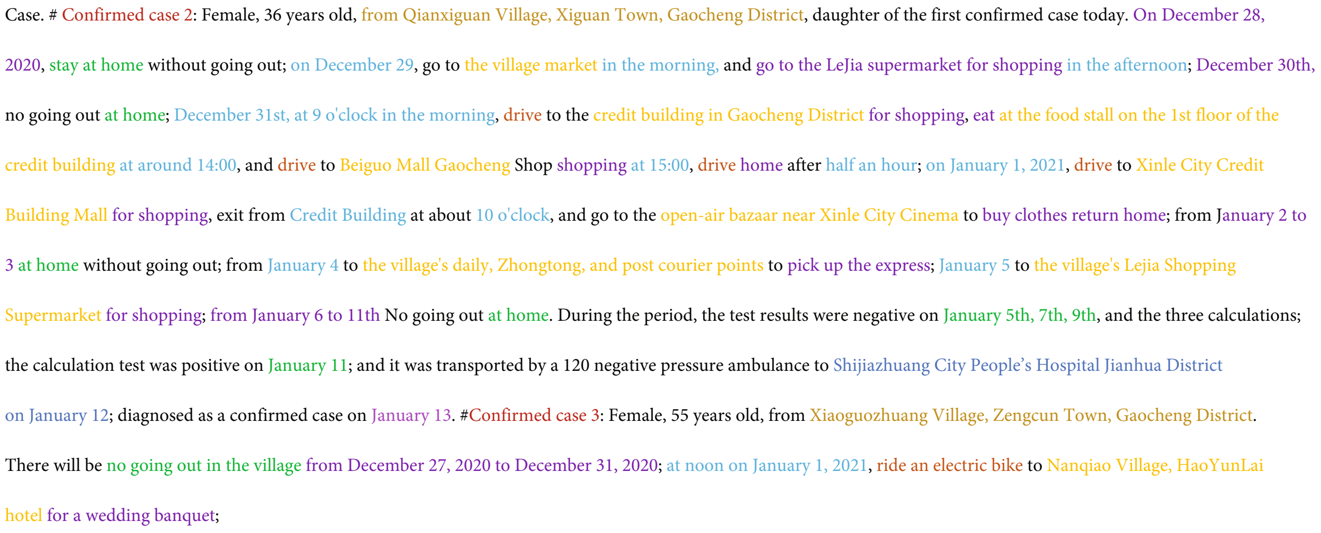
1. Introduction

In December 2019, a novel coronavirus (COVID-19) was transmitted between species and spread fast worldwide in a short period. The fast spread of the disease and severe economic and social devastation are far beyond people's expectations. Utilizing data mining and natural language processing technology to anticipate epidemic development trends and carry out intelligent security early warning has piqued the interest of academics. It has progressively become one of the hotspots in natural language processing [1–3]. According to the current standardization of epidemic prevention and control, the epidemic spread is characterized by random small-scale bursts, which leads to a shortage of case information data in the early stage of transmission, making the study of epidemic information more complex and problematic. With their high efficiency, standardization, and real-time features, artificial intelligence, big data, and other technologies have achieved tremendous gains in several epidemic prevention and control domains since the outbreak of novel coronavirus-infected pneumonia [4–6]. The SEIR model is one of the most extensively used epidemic prediction models in the epidemic dynamics model. It may consider transmission speed and mode and numerous infectious disease prevention and control strategies. The trajectory location early warning model monitors and provides an early warning by mining semantic trajectory data and combining location, time, and various application circumstances. Furthermore, some researchers use multivariate cosmos algorithms with artificial intelligence algorithms, natural language processing technology, and other comprehensive considerations of different prevention and control measures and various factors to build an early warning model, which has obvious advantages. COVID-19 pandemic prediction has piqued the interest of academics from many sectors all around the world. For example, Khayyat et al. [7] provided a predictive analytic model to anticipate the spread of the epidemic in Saudi Arabia and utilized the time series correlation FB model to perform a t-test on the data, giving a foundation for the epidemic's future development. Alsunaidi et al. [8] established a dynamic prediction and transmission technique based on the SEIRD epidemic model and calculated time-varying model parameters using maximum likelihood to show the emotional influence of the infection rate, death rate, and recovery rate on COVID-19 transmission. Balaha et al. [9] extracted characteristics from CT scans and learned them using deep learning and pretrained models. A hybrid technique that superimposes several CNN models is adopted to improve prediction accuracy. Wieczorek et al. [10] developed a neural network model for epidemic spread utilizing government data and the Nadam training model to achieve good prediction accuracy; Kozio et al. [11] offered a fractional-order SIR epidemic model to forecast epidemic spread. The model parameters were estimated using the genetic method, and the simulation was validated using Spanish data. Because the critical content of the text information of the epidemic has been centralized, each phrase has its core subject, such as the basic information about the infected patients, the track information during the sickness, and the time of diagnosis. When conducting information extraction, the information extraction technology indicated in the current results frequently cannot take advantage of the peculiarities of the epidemic notice text and cannot use the topic information. Furthermore, some algorithms suffer high labeling costs and little labeled data. Due to the abovementioned issues, this work presents a cooperative learning strategy for topic recognition and named entity identification TBR-NER. This technique annotates the epidemic notification information first and then exploits the features of distinct entities belonging to different subjects to accomplish topic self-annotation without adding extra annotation data, decreasing human annotation effort.

To increase the model's generalization ability in the absence of samples, entity mention replacement and external knowledge base replacement are employed to improve data. At the same time, transfer learning enhances the model's text parsing capacity. The following are the primary contributions of this study. (1) A TBR-NER model is presented for text extraction of epidemic notification information. The text extraction accuracy of epidemic notification information is increased by integrating the notification information text training mode with topic recognition in entity recognition. (2) Propose a self-annotation approach for cooperative learning. Topic self-annotation is done without providing new annotation data based on entity annotation. The features that distinct entities belong to various topics are utilized for topic recognition jobs. (3) Add the Chinese information corpus (Chinese Wikipedia corpus) to the training session. On the one hand, some information about entity place names is replaced by entities during the training process to achieve the goal of data enhancement. On the other hand, the pretrained model is finally transferred to the joint learning model by learning the notice text to express the learning sentence structure, word collocation information, and so on, which effectively avoids the diversity and ambiguity of the expression and improves the model's generalization ability. (4) Comparative tests were conducted on the collected datasets from the provinces of Jilin and Hebei. The findings demonstrate that the proposed topic recognition and named entity recognition (TBR-NER) combined learning system can extract critical information in the epidemic notification job. The topic recognition and entity recognition tasks will help in the cooperative learning process, resulting in higher classification results.

2. Related Work

2.1. Conditional Random Field. The conditional random field (CRF) is a probabilistic model [12] used to label and split data with a sequence structure. It combines the properties of the maximum entropy model with the hidden Markov model, and it can describe long-distance dependence. It can globally normalize the characteristics and then achieve the objective of global optimization, which better overcomes the label bias problem. When first collecting text information, the model typically pulls crucial information in the form of the character level. Even though this approach is basic and straightforward to master, it has certain flaws. Discerning things based on categorization labels is complicated when multiple continuums coexist. For example, the three geographical names of Jilin province, Changchun city, and Jilin University correspond to three-place entities in the label categorization of Jilin University in Changchun city, Jilin province. However, there is no visible border between continuous entities when using the character-level form to extract modeling and the abovementioned words will be predicted as an entity label. To address this issue, this article adds the entity label BIO (see Figure 1). Each original independent entity information annotation is labeled with one of three types of labels: the beginning label (B-), the intermediate label (I-), and the irrelevant label (O-). See Figure 1. The classification prediction results of "Qianxiguan village, Xiguan town, Gaocheng district, 'become' B-location, I-location, I-location, B-location, I-location, I-location, B-location, I-location, I-location, I-location, I-location, and I-location." The red label represents the patient's ID, the dark-green label represents the patient's hometown, the pink label represents the patient's home time, the purple label represents the patient's trajectory action event, the blue label represents the patient's opening trajectory movement time, and the light-green label represents the patient's diagnosis time; the dark-blue label represents the time when the



Case. # Confirmed case 2: Female, 36 years old, from Qianxiguan Village, Xiguan Town, Gaocheng District, daughter of the first confirmed case today. On December 28, 2020, stay at home without going out; on December 29, go to the village market in the morning, and go to the LeJia supermarket for shopping in the afternoon; December 30th, no going out at home; December 31st, at 9 o'clock in the morning, drive to the credit building in Gaocheng District for shopping, eat at the food stall on the 1st floor of the credit building at around 14:00, and drive to Beiguo Mall Gaocheng Shop shopping at 15:00, drive home after half an hour; on January 1, 2021, drive to Xinle City Credit Building Mall for shopping, exit from Credit Building at about 10 o'clock, and go to the open-air bazaar near Xinle City Cinema to buy clothes return home; from January 2 to 3 at home without going out; from January 4 to the village's daily, Zhongtong, and post courier points to pick up the express; January 5 to the village's Lejia Shopping Supermarket for shopping; from January 6 to 11th No going out at home. During the period, the test results were negative on January 5th, 7th, 9th, and the three calculations; the calculation test was positive on January 11; and it was transported by a 120 negative pressure ambulance to Shijiazhuang City People's Hospital Jianhua District on January 12; diagnosed as a confirmed case on January 13. #Confirmed case 3: Female, 55 years old, from Xiaoguozhuang Village, Zengcun Town, Gaocheng District. There will be no going out in the village from December 27, 2020 to December 31, 2020; at noon on January 1, 2021, ride an electric bike to Nanqiao Village, HaoYunLai hotel for a wedding banquet;

Figure 1: NER entity boundary tag BIO model display diagram.

patient was transferred, the lavender label represents the name of the hospital where the patient was transferred, and the dark-brown label represents the patient's transportation information.

Although adding BIO can help the model different continuous things, this strategy will have some drawbacks. For example, as the number of labels rises, the danger of an unlawful prediction label sequence increases. The following faults are possible in this paper's epidemic identification task: (1) The model recognizes work information labels in the case of Id, such as "B-case id, I-work information, and I-case Id." (2) The initial label "B-case Id" should appear at the beginning of entity information, and the model forecasts in the middle of entity information, such as "I-case id, B-case id, and I-case id." These incorrect labels will have an impact on the model's overall accuracy. As a result, in label prediction, it is necessary to examine the word-level classification prediction accuracy as much as possible to assure the prediction of the lawful label sequence.

The introduction of the conditional random field model (CRF) in this study may identify the part of the label sequence distribution rules, allowing the label sequence to be legalized; the formula is as follows:

$$\text{Score}(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1}), \quad (1)$$

$$p(l|s) = \frac{\exp\left[\text{score}(l|s)\right]}{\sum_{l'} \exp\left[\text{score}\left(l'|s\right)\right]}. \quad (2)$$

Among them, formula (1) represents the score of the feature function on the label sequence I, s represents the sentence that needs to be labeled with a part of speech, i represents the ith word in sentence s, l_i represents the part of speech labeled by the label sequence to be scored for the i th word, and l_{i-1} indicates the labeling sequence that needs to be cut to mark the position of speech for the $i-1$th term. The outer summation is the sum of the core values of each feature function f_j, and the inner outline is the sum of the feature values of the words at each position in the annotated sentence. Formula (2) is to index and standardize the score, and the probability value $p(l|s)$ of the label sequence I can be obtained. In this paper, the Viterbi algorithm [13] is used to calculate the score of legal label paths quickly and dynamic programming DP is used to solve the optimal path problem in this paper.

2.2. Transfer Learning. Transfer learning [14, 15] is characterized by a given labeled domain and an unlabeled target domain. Even though the data distribution in the two parts is different, the knowledge of D_t can be learned through the knowledge of D_s. Transfer learning is an optimization approach introduced for solving problems with a small number of samples. Transfer learning is typically appropriate when the amount of data in the source domain is sufficient and the amount in the target domain is minimal.

This work uses transfer learning to aid model training to address the underfitting problem caused by a small amount of epidemic information data. Because the quantity of notification messages is frequently insufficient in the early stages of epidemic transmission, only a small amount of data is used for model training, resulting in the undergeneralization of the model. As a result, this research gathers a Chinese information notification text corpus from Wikipedia. It employs a BERT model based on mask language (MLM) for self-supervised training to gain additional text collocation and sentence structure information.

The Bidirectional Encoder Representations from Transformer (BERT) language preprocessing paradigm can automatically extract rich word-level characteristics, grammatical and structural features, and semantic information from sequences. The BERT model utilized in this study includes 12 layers, and the output of each hidden layer after pretraining may convey word vectors to varying degrees. The hidden layer output vector of the last layer is commonly used in the BERT model, although some tests demonstrate that the later the layer vector is, the more difficult it is to match the present job itself [16]. To improve the retention of high-dimensional abstract information and the generalization ability of the pretraining model, this paper fuses the outputs of the last four layers of the pretrained model $k_1 * a_{nt} + k_2 * a_{(n-1)t} + k_3 * a_{(n-2)t} + k_4 * a_{(n-3)t}$, where the a_{nt} is the model output vector, n is the number of layers, t is the time step, and the weight coefficients $k1$, $k2$, $k3$, and $k4$ are 0.1, 0.2, 0.3, and 0.4, respectively.

2.3. Rule Matching. Rule matching is a way of searching for matches in highly regularized data [17]. When extracting epidemic information in this research, it is discovered that some data information is very standardized, as shown in Figure 2. For example, a "certain X-year-old" statement has a greater confidence that the statement describes the age or a "XXXX year was born" statement has higher confidence that the information describes the date of birth.

The examination of the epidemic text reveals that several fundamental attribute interactions exhibit consistent patterns in the text. By evaluating the samples and removing interference factors such as manual annotation mistakes, the regular expression is utilized to extract the information attribute connection. The fusion data is searched using matching. Using the rule matching strategy to lower the cost of label annotation might focus the model on less regular label categorization, improving the model's accuracy.

2.4. Data Enhancement. Data augmentation is a technique that synthesizes new data from existing data. When training samples are sparse and labeling costs are high, it seeks to enhance the number of data points, reduce overfitting, and improve the model's overall generalization ability [18, 19].

Due to a lack of case texts in the epidemic's early stages, the model struggles to learn more detailed rules based on a few samples. Furthermore, several tags in the pandemic text, such as transit information and employment information, have a substantially lower number of occurrences than other tags. Without data augmentation, the model cannot detect the basic information of such labels. Synonym substitution (SR), random insertion (RI), spontaneous exchange (RS), random deletion (RD), and entity mention substitution are all typical data augmentation strategies used in NLP jobs [20]. The methods of synonym substitution and entity mention substitution are primarily used in this article. However, using synonymous substitution (SR) may not have the desired impact of fundamentally improving data. This is because the vector values of the interchangeable substitute terms are almost identical, resulting in enhanced subjective data. On the other hand, the trained model may perceive the augmented data as the exact phrase and there is no significant data extension.

This paper presents a solution to the above-listed challenges: first, the part of speech of each word in the original text is identified, next, comments with the same amount of speech but different are disrupted into other sentences to replace synonyms, and finally, the credibility of the data-enhanced text is checked and the data-enhanced text is merged into the training corpus, for example, "Dining at an HSBC Hotel" and "Shopping in Joy City on X," where words like "HSBC Hotel" and "Joy City" can be identified as synonymous with different words; "Dining" and "Shopping" belong to the same entity type, and the enhanced data "dining in the Grand Hyatt" can be obtained through the improved synonym substitution scheme.

3. Construction of the Joint Learning Model Based on Topic Recognition and Named Entity Recognition

3.1. Principle of the TBR-NER Model. The corpus required for the TBR-NER joint learning model comes from two aspects: (1) the new epidemic notification corpus was extracted according to the epidemic-related information announcement issued by the National Health Commission and (2) the Chinese notification corpus was obtained from the Chinese Wikipedia corpus. After removing the text information, the two parts of the canon are data cleaned and the epidemic notification corpus is annotated. Then, the outbreak notification corpus is data augmented using entity mention replacement and external knowledge base replacement. Secondly, pretrain the Chinese information notification text corpus in Wikipedia using the MLM language model to construct the corresponding dictionary mapping. At the same time, after data enhancement of the epidemic corpus, the related word vector needs to be obtained according to the index. Thirdly, a standard learning method based on topic recognition and named entity recognition (TBR-NER) is used to train and predict the marked information. The effect of entity recognition and classification is improved with the help of the topic of text information. Finally, after the postcorrection of the results, the optimal results are obtained. The TBR-NER model framework is shown in Figure 3.

3.2. BERT Model Based on Masking Language (MLM). The BERT model based on MLM is a new language representation model released by the Google AI team in 2018, which is the bidirectional encoder representation of Transformer [21]. Unlike other language representation models, BERT is aimed at pretraining a deep bidirectional representation by jointly adjusting the context in all layers. Thus, the pretrained BERT representation can be fine-tuned by an additional output layer. It does not need to modify the architecture of specific tasks and can be applied to various fields of task model construction.

There are two commonly used BERT models:

(1) BERTBASE: $L = 12$, $H = 768$, $A = 12$, and total parameter = 110 M

(2) BERTLARGE: $L = 24$, $H = 1024$, $A = 16$, and total parameter = 340 M

BERT has excelled in several NLP tasks, including categorization, question answering, and translation [22–24]. A 12-layer BERT model is used in this work. The model's

#无症状感染者2：男，1978年出生，通化市人，住址为通化市东昌区东苑小区，系1月15日通报的无症状感染者6的儿子。1月

#Asymptomatic infected person 2: Male, born in 1978, a native of Tonghua City, whose address is Dongyuan Community, Dongchang District, Tonghua City. He is the son of asymptomatic infected person 6 reported on January 15. January

#无症状感染者3：男，1955年出生，通化市如人，住址为通化市东昌区厚德载物C区，系1月17日通报的无症状感染者21的丈夫。

#Asymptomatic infected person3: Male, born in 1955 in Tonghua City, where he lives in Zone C, Houdezaiwu, Dongchang District, Tonghua City. He is the husband of asymptomatic infected person 21 notified on January 17.

Figure 2: Have highly normalized data.

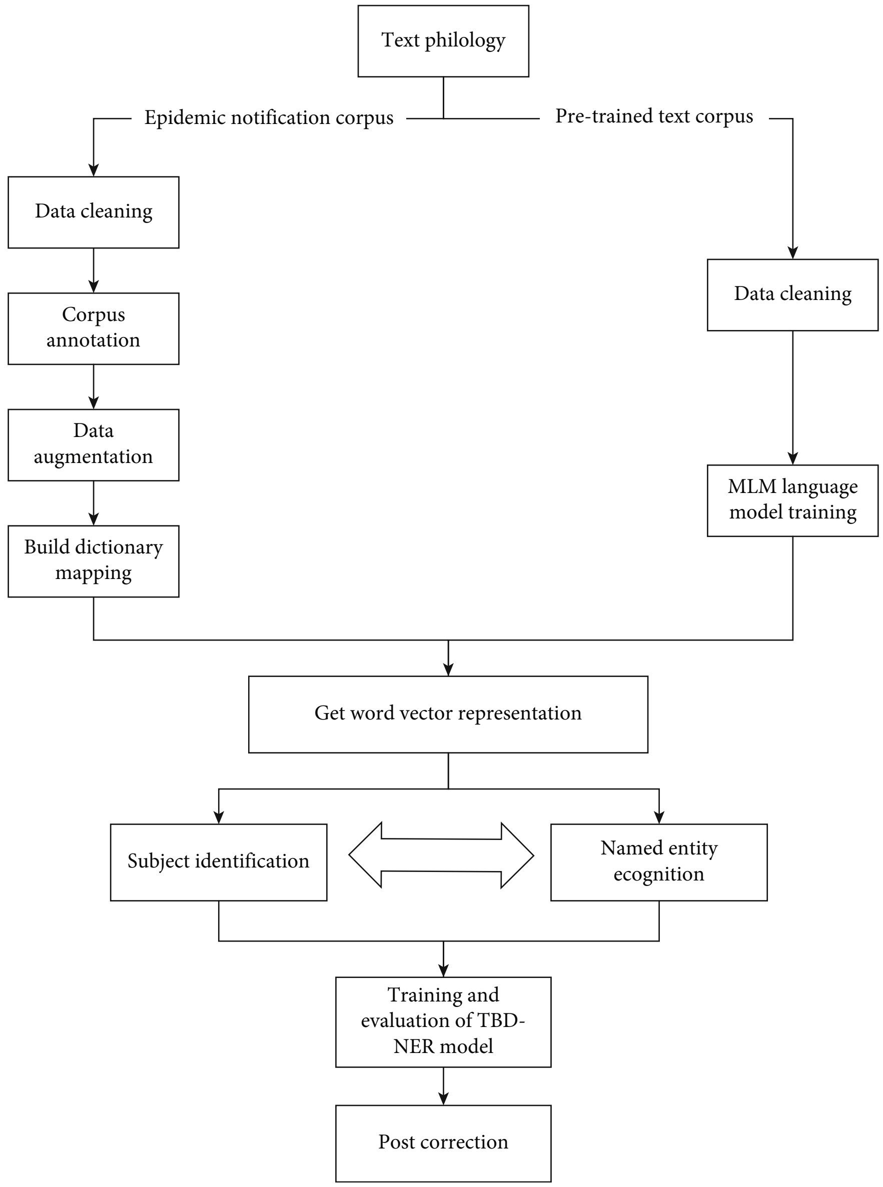


Figure 3: Flow chart of model construction.

primary distinction from prior models is that it recommends predicting the LOSS value of the following phrase by transforming a limited number of words into masks or randomly changing another word with a specific probability. The primary purpose of the BERT model is to acquire word vectors via massive text learning, assess the relationship between distinct token units, and then map the association to variables. The BERT model employs two-stage model training to address the widespread occurrence of polysemy in the notification information. It starts with two-way language model training and then moves to the fine-tuning mode to address downstream objectives. As a result, the word vector taught by BERT has a high impact and flexibility.

3.3. Two-Stage Hierarchical Learning Model Based on Topic Prediction. For epidemic text information, first divide each epidemic notification text into five topics: case basic information introduction, case trajectory information during illness, home information during case illness, case diagnosis information, and other types. Then, the entity information is divided into 16 categories: patient ID, the patient's native place, the patient's residence, the patient's working class, the patient's diagnosis time, the patient's opening trajectory movement time, the patient's moving trajectory starting place, the patient's moving trajectory termination place, the patient's trajectory action event, the patient riding the transportation tool, the transportation tool information, the patient's home time, the patient's diagnosis time, the patient being transported by the hospital time, the patient being fascinated by the hospital name, and the irrelevant information. The 15-type entity information corresponds to BI's start and intermediate labels in sequence annotation and irrelevant entity information. As a result, 15 types of entity information matching 30 category labels, plus unrelated classes, must be forecasted for each letter as a category in 31 categories.

The two-stage hierarchical learning concept is as follows to improve the model's prediction of label information: first, the text statement is used to categorize the topics. Following completion of the categorization, the entity classification is based on the labels that may exist in the current statement topic. Figure 4 depicts a two-stage hierarchical learning model based on topic prediction. Two-stage hierarchical learning is utilized to scatter the titles of the original 31 categories into multiple topic categories. The label types for each topic are few, and the model is difficult to misinterpret.

Two-stage hierarchical learning based on topic prediction is mainly divided into two steps:

(1) Subject classification is carried out on a notification information text, and the text information is divided into one of five types of subjects

(2) For example, the text of the epidemic information notification is as follows in "case 1: female, 18 years old, from Shijiazhuang, Hebei, salesperson of Ping An supermarket, and living in Ping An community. On January 18, she ate at HSBC Hotel. She did not get out from home on January 19. On January 20, the nucleic acid test was positive. On January 21, after consultation with the expert group, it was diagnosed as new coronary pneumonia." In the abovementioned epidemic notification information, identifying "Ping An community" as a "patient residence" requires two steps: first, it is necessary to determine its topic. The sentence is as follows in "case 1: female, 18 years old, from Shijiazhuang, Hebei, a salesperson in a Ping An supermarket, and living in a Ping An community." It belongs to the topic category "introduction to basic information of cases;" then, under the topic of "introduction to basic information of cases," use model 1 to identify and classify entities. The possible entity categories under this topic include patient id, patient origin, patient residence, patient workplace, and irrelevant information

3.4. Two-Stage Joint Learning Based on Topic Prediction. The suggested two-stage hierarchical learning based on topic prediction used the subject information of the epidemic notice text in the last part. However, there are still the following difficulties in the actual application scenario:

(1) For error accumulation problem due to entity label prediction after two stages, if the topic prediction model is weak in the first step, the total results are not very good no matter how accurate the succeeding entity recognition model is. The subject prediction error accumulates in the first step, and the influence of the subject prediction becomes the bottleneck of the succeeding prediction

(2) For the problem of model redundancy, in two-stage hierarchical learning, a large number of models need to be used for different tasks, which is more cumbersome in practical applications

(3) Splitting the two tasks of subject information and entity recognition of sentences in the task splitting problem prevents the two pieces of information from supplementing each other. The improvement in the entity recognition effect in the second stage will have no feedback effect on the sentence classification effect. Improved topic classification accuracy will simply enhance the bottleneck of the total classification impact, and there is no benign mutual reference between topic recognition and entity recognition

Based on the abovementioned issues, this work suggests a novel two-stage joint learning technique based on topic prediction: topic-based discrimination. TBR-NER is based on the combined learning model of topic recognition and named entity recognition. TBR-NER is a step forward based on the hierarchical learning paradigm. The original topic recognition and entity recognition are fused based on the exact usage of the subject information of the notification text, allowing the model to fulfill the two tasks of topic recognition and entity recognition simultaneously, as illustrated in Figure 5.

The TBR-NER model proposed in this paper is divided into the following steps:

(1) First, perform character-level segmentation on each sentence. This article does not use the conventional method of first-word segmentation, and then, input it into the model because word segmentation will have a priori bias. Different word segmentation algorithms will aggregate other numbers of characters into words, and these presegmented words are not necessarily suitable for the current task. BERT's

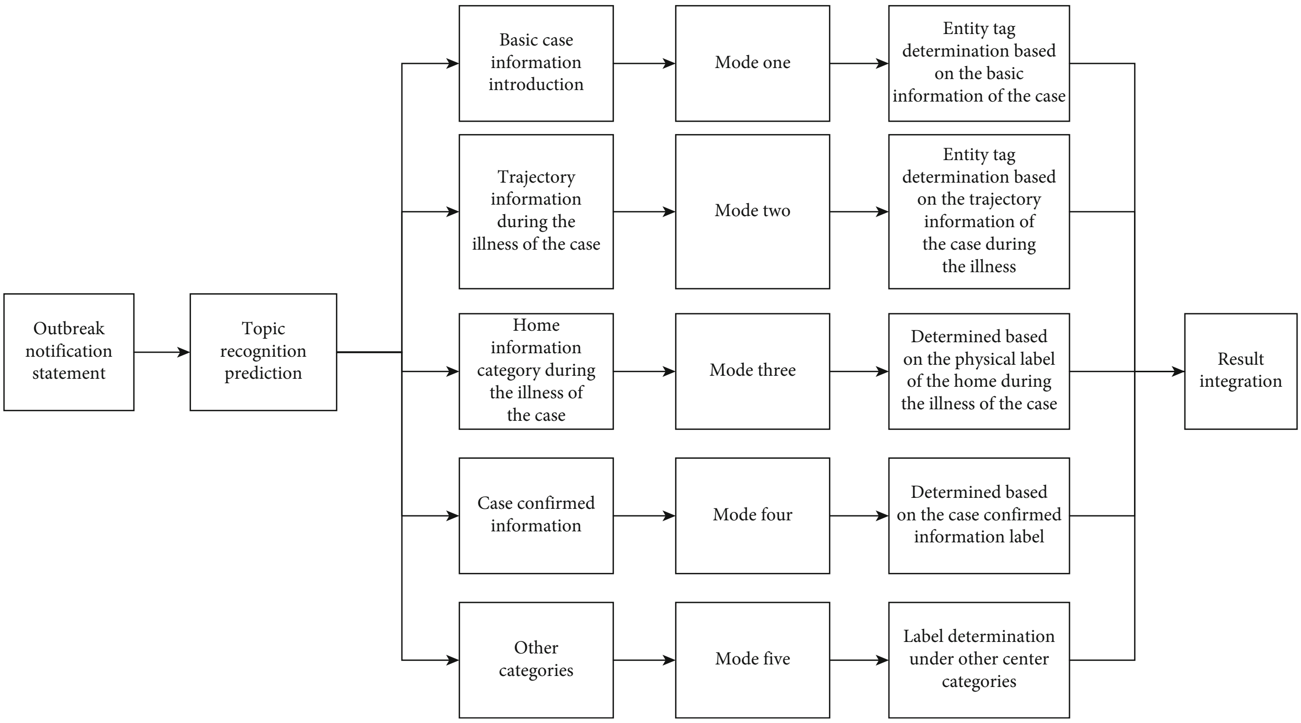

Outbreak notification statement
Topic recognition prediction
Basic case information introduction
Mode one
Entity tag determination based on the basic information of the case
Trajectory information during the illness of the case
Mode two
Entity tag determination based on the trajectory information of the case during the illness
Home information category during the illness of the case
Mode three
Determined based on the physical label of the home during the illness of the case
Case confirmed information
Mode four
Determined based on the case confirmed information label
Other categories
Mode five
Label determination under other center categories
Result integration

Figure 4: A two-stage hierarchical learning model based on topic prediction.

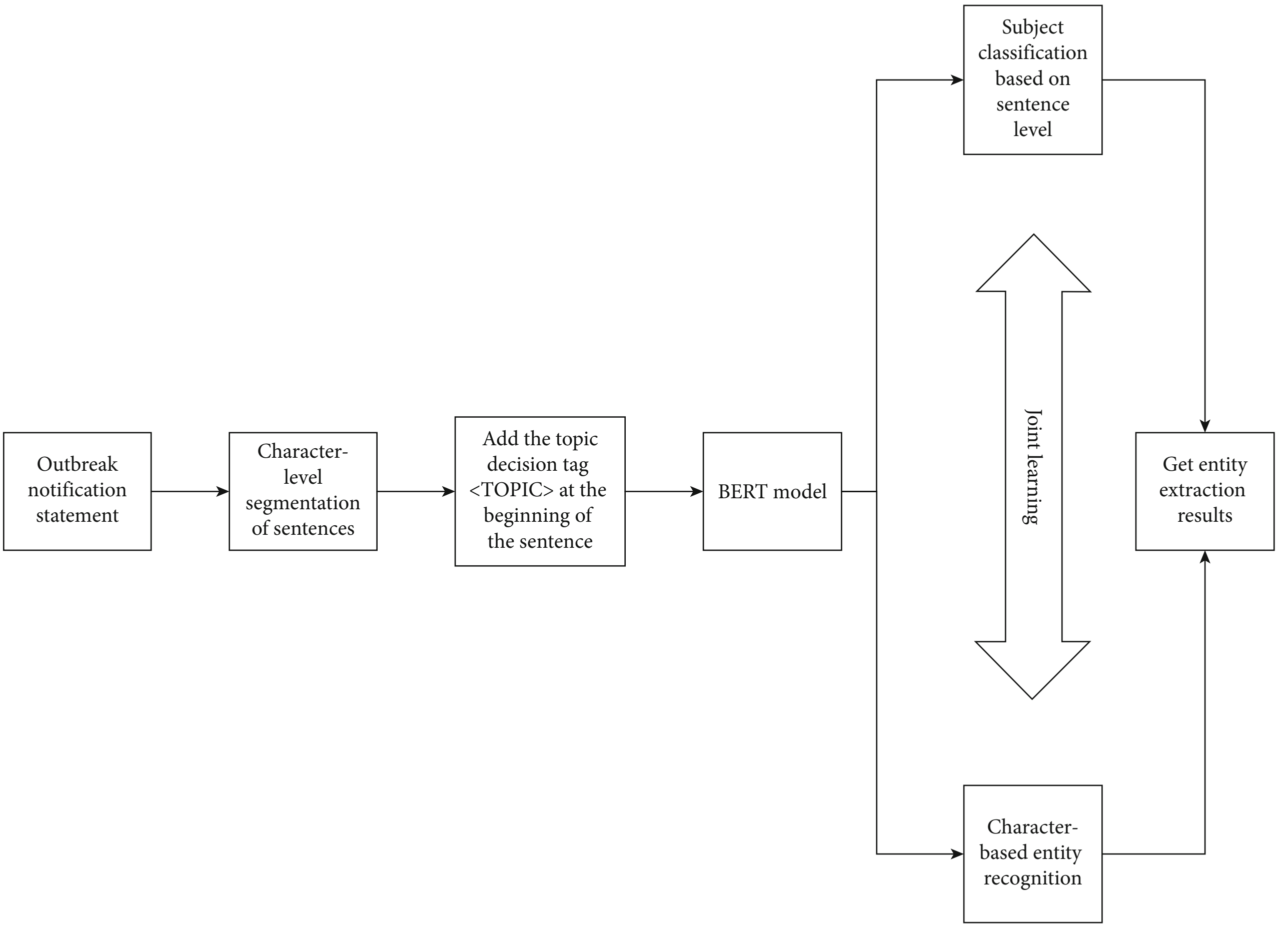

Outbreak notification statement
Character-level segmentation of sentences
Add the topic decision tag <TOPIC> at the beginning of the sentence
BERT model
Subject classification based on sentence level
Joint learning
Character-based entity recognition
Get entity extraction results

Figure 5: TBR-NRE model.

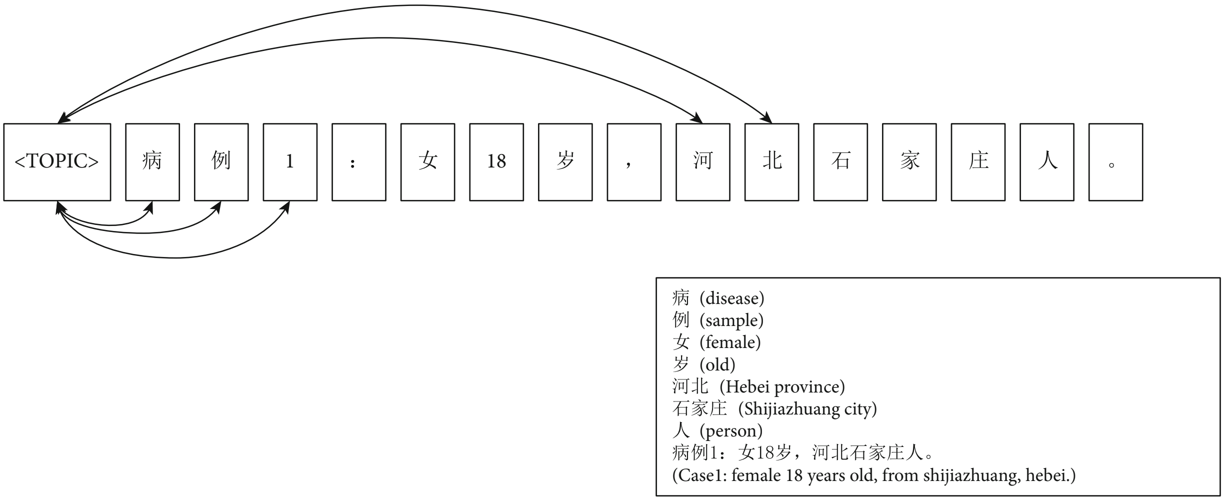

<TOPIC>
病
例
1
：
女
18
岁
，
河
北
石
家
庄
人
。
病 (disease)
例 (sample)
女 (female)
岁 (old)
河北 (Hebei province)
石家庄 (Shijiazhuang city)
人 (person)
病例1：女18岁，河北石家庄人。
(Case1: female 18 years old, from shijiazhuang, hebei.)

Figure 6: The mutual learning process of topic recognition and entity recognition.

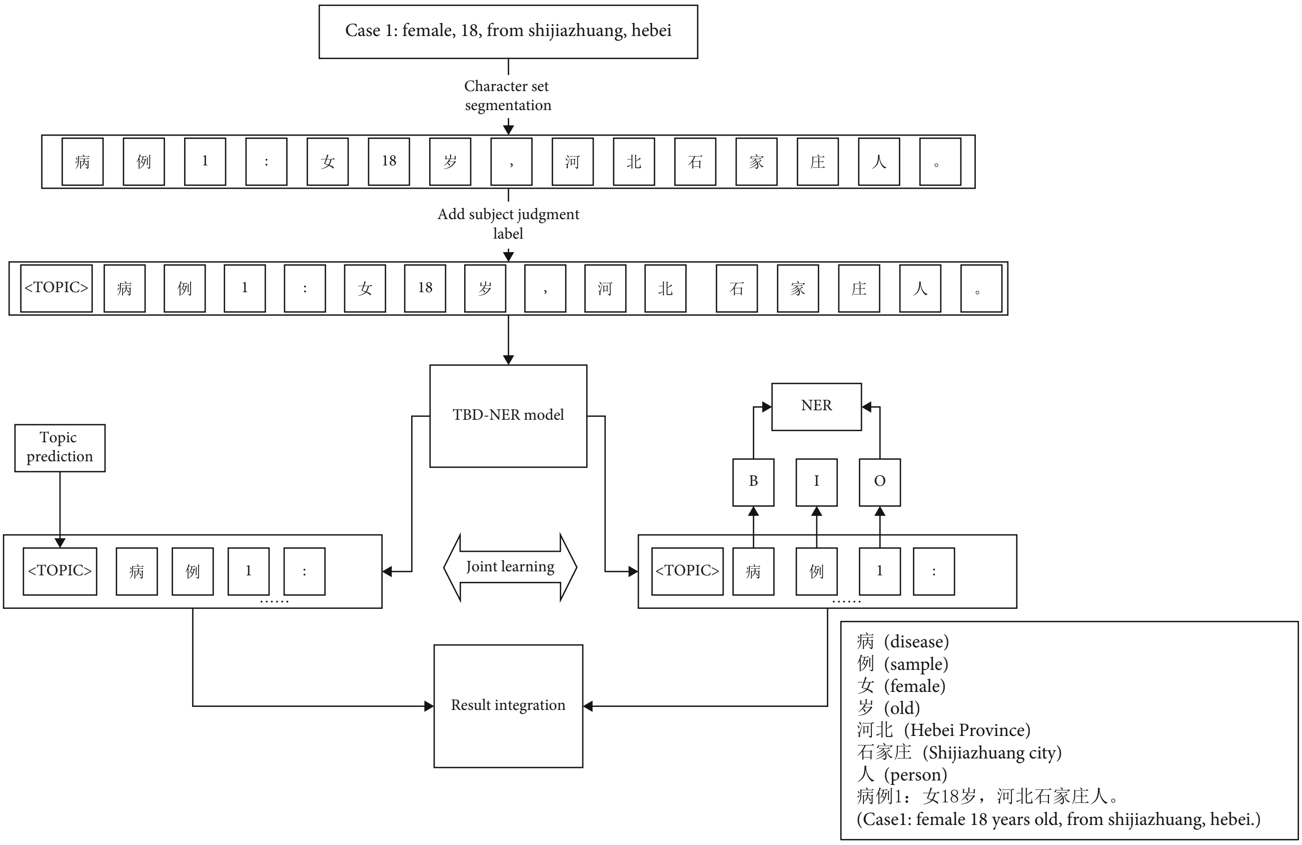

Case 1: female, 18, from shijiazhuang, hebei
Character set segmentation
病 例 1 ： 女 18 岁 ， 河 北 石 家 庄 人 。
Add subject judgment label
<TOPIC> 病 例 1 ： 女 18 岁 ， 河 北 石 家 庄 人 。
TBD-NER model
Topic prediction
<TOPIC> 病 例 1 ：
......
Joint learning
NER
B
I
O
<TOPIC> 病 例 1 ：
......
Result integration
病 (disease)
例 (sample)
女 (female)
岁 (old)
河北 (Hebei Province)
石家庄 (Shijiazhuang city)
人 (person)
病例1：女18岁，河北石家庄人。
(Case1: female 18 years old, from shijiazhuang, hebei.)

Figure 7: TBR-NER model learning case.

self-attention mechanism model will aggregate characters according to the current job at the low level. Therefore, this paper chooses Chinese characters or numbers as the split unit

(2) For topic self-labeling based on entity labeling results, add a "<TOPIC>" tag to the beginning of each sentence. This tag is used for the topic classification of the current correction, which contains the topic information of the current sentence, as shown in Figure 6. The topic information in this article does not require additional artificial annotation. The topic is determined by the content of the entity tag marked in the previous stage, which is a "self-labeling" process. For example, when a "case id" or "case origin" entity appears in a sentence, the sentence will be automatically marked as a "case basic information introduction category." The topic self-labeling

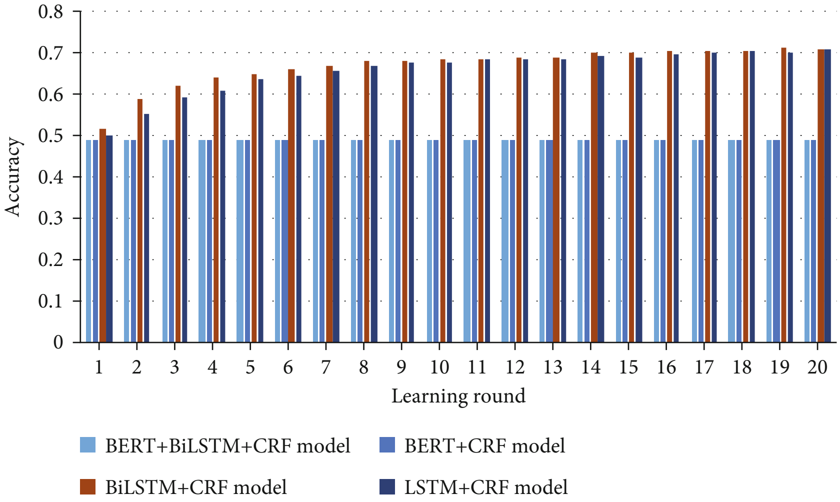


FIGURE 8: Four kinds of NER model entity recognition independent test set accuracy.

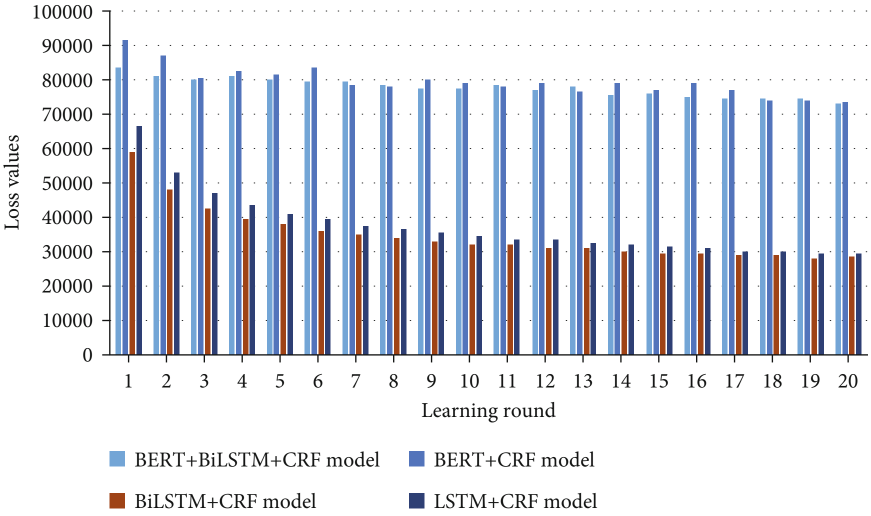


FIGURE 9: Loss values of four NER models.

process based on entity labeling results saves the manual labeling process and does not introduce additional knowledge

(3) Predict the topic category of the sentence where the "<TOPIC>" tag is located, and perform entity recognition on the sentence content. Since "<TOPIC>" itself is located in the sentence, the model can also observe the topic representation of "<TOPIC>" while doing entity recognition and "<TOPIC>" itself can also be used for topic classification. See the entity representation of the different characters themselves. Topic recognition and entity recognition learn from each other to achieve better prediction results. The learning process of the TBR-NER model is shown in Figure 7

3.5. TBR-NER Model Loss Function. The model loss function of TBR-NER proposed in this paper mainly consists of the following two parts:

$$\text{Loss} = \partial \text{loss}_{\text{idea}} + \beta \text{loss}_{\text{ner}}. \quad (3)$$

∂ and β correspond to their proportion weights, and by adjusting them, the model can pay more attention to a certain task. $\text{Loss}_{\text{idea}}$ represents the prediction loss of a sentence, which is called cross-entropy, and the formula is as follows:

$$\text{Loss}_{\text{idea}} = -\sum_{x} p(x) \log q(x). \quad (4)$$

Among them, $p(x)$ is the actual probability distribution and $q(x)$ represents the predicted probability distribution.

Loss_{ner} is the loss function of the conditional random field (CRF) optimization algorithm introduced in the summary of Section 3.1, and the formula is as follows:

Table 1: Analysis and comparison of four classical NER models.

	BERT + BiLSTM + CRF	BERT + CRF	BiLSTM + CRF	LSTM + CRF
Accuracy rate	0.4904	0.4904	0.7132	0.7098
Loss value	73349.1 9	73936.48	28270.13	29563.14

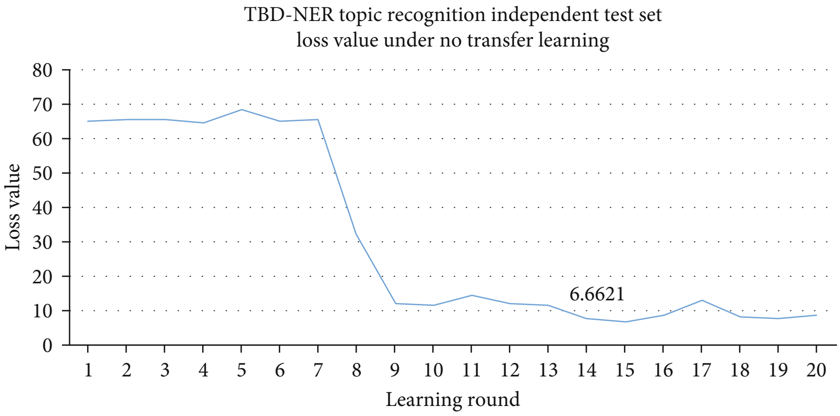


Figure 10: TBR-NER-independent test set accuracy without transfer learning.

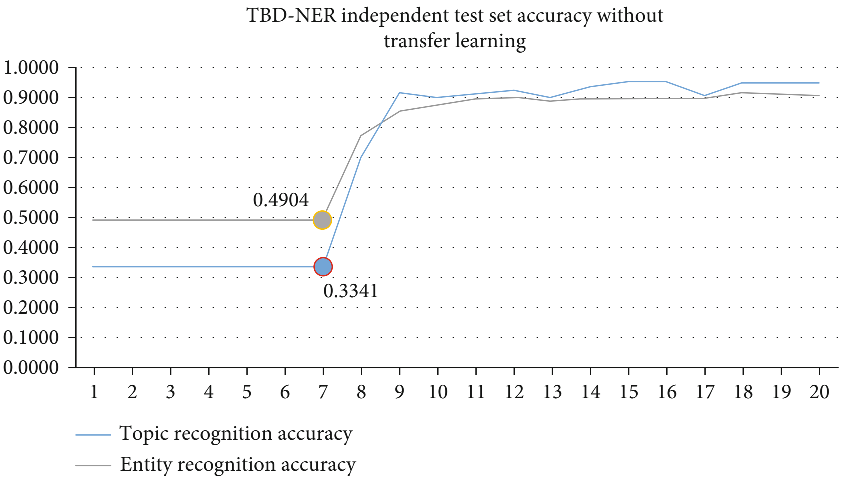


Figure 11: TBR-NER topic recognition with independent test set loss values without transfer learning.

$$\mathrm{Loss}_{\mathrm{ner}} = p(y_1, y_2, \cdots, y_n|x) = \frac{1}{z(x)} e^{h(y_1:x)+\sum_{i=2}^{n} g(y_i : y_{i-1})+h(y_i : x)} . \tag{5}$$

Among them, $h(y_i : X)$ represents the emission fraction, $g(y_i ; y_{i-1})$ represents the transfer score, $z(x)$ represents the normalization factor, and $P(y_1, y_2, \cdots, y_n|X)$ mark the probability of $y_1, y_2, \cdots, y_n$.

4. Experimental Results and Analysis

4.1. Analysis of Experimental Results. The text information of epidemic notifications issued by the Health Commission in Jilin province and Hebei province from January 10 to February 14, 2021, was chosen as the data source to test the validity of the suggested approach in the task of epidemic notification information. The dataset was chosen randomly, with 70% as a training set and 30% as an independent test set. The accuracy and ultimate loss of the separate test set are utilized as evaluation model indicators, with accuracy defined as follows:

$$\mathrm{ACC}_{\mathrm{word}} = \frac{\sum \mathrm{token}_i}{\mathrm{sum(word)}} . \tag{6}$$

In formula (6), sum(word) represents the sum of all characters and token_i represents the sum of correctly predicted characters, which is the accuracy of the expected characters. In this study, four NER models were selected as the comparative reference experiments: BERT + BiLSTM

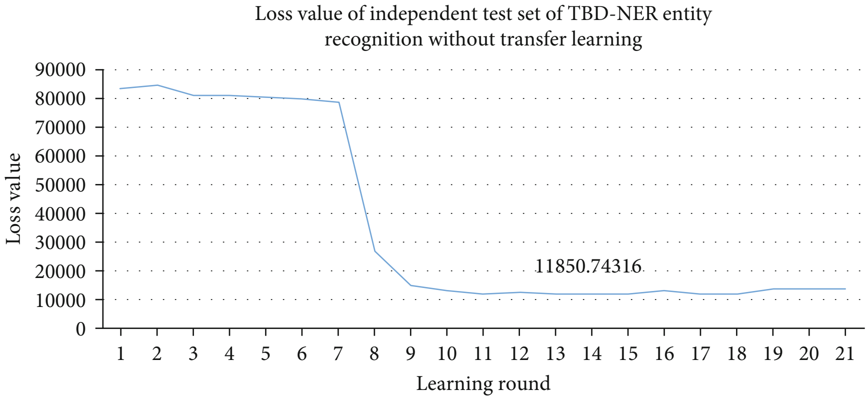


FIGURE 12: Loss values of independent test sets for the TBR-NER entity.

+ CRF [25], BERT + CRF [26], BiLSTM + CRF [27], and LSTM + CRF [28]. The experimental results of each model's first 20 rounds of learning were compared.

4.2. Model without Transfer Learning. The accuracy and loss values of four NER models in entity recognition are compared without transfer learning. The accuracy rates of the four NER models are relatively low, as seen in Figures 8 and 9. In 20 rounds of education, the BERT + BiLSTM + CRF model and the BERT + CRF model were both less than 0.49. They cannot converge for an extended period throughout the training process, resulting in gradient dispersion. The model cannot break past the bottleneck, resulting in a significant loss of values and an optimal local dilemma. Although the BiLSTM + CRF and LSTM + CRF models outperform the previous two models and the model loss values converge fast, the model accuracy remains poor. Table 1 compares the accuracy and loss values of the NER model for entity recognition in 4.

To assess the effect of entity recognition fairly, the TBR-NER model also performs 20 rounds of learning without incorporating transfer learning and makes entity recognition prediction and topic recognition prediction using epidemic trajectory information. The experimental findings are depicted in Figures 10–12. The TBR-NER model's fitting performance has been dramatically enhanced compared to that of the previous four NER models. Under the same settings, the accuracy of the TBR-NER model after the seventh cycle of learning. Table 2 displays the accuracy and loss values of topic recognition tasks and entity recognition tasks using TBR-NER without transfer learning.

The experimental findings show that the TBR-NER model can overcome the loss bottleneck after many learning rounds and predict with high accuracy. This is because the model entity recognition task relies heavily on the topic categorization job. The TBR-NER model's main algorithm is a two-stage joint learning technique based on topic recognition. The topic classification task of the statement is straightforward in the iterative stage of the initial model, and the topic representation vector has more excellent separability in high latitudes (see Figure 13). The entity recognition task is driven by the significant growth in the subject categorization job. Topic recognition and entity recognition continue to learn in the following learning phase, achieving synchronous convergence. The model's accuracy under entity recognition and classification reached the superior value of 91.60 percent when the TBR-NER model was taught in the 18th round. Although the TBR-NER model considerably improves prediction outcomes when compared to its four types of NER models, the model's accuracy does not achieve the optimum state in the first seven learning periods. After the model passes through the loss domain, its lost value rapidly converges and the model's accuracy eventually tends to be stable.

TABLE 2: Analysis and comparison of the TBR-NER model without transfer learning.

	TBR-NER model for topic recognition	TBR-NER model entity recognition
Accuracy rate	0.9709	0.9160
Loss value	6.0467	11850.74

4.3. Model under Transfer Learning. This work employs the transfer learning auxiliary model to train and improve the model's generalization performance to address the underfitting problem caused by the small amount of epidemic information data. To maintain the fairness of the control experiment in the prior summary experiment, the abovementioned TBR-NER model does not apply the transfer learning optimization model. In this overview, transfer learning will optimize the TBR-NER model and compare the outcomes.

Figures 14–16 exhibit the accuracy of topic recognition, entity recognition, and loss value of the TBR-NER model under transfer learning. When the TBR-NER model learns in the sixth round, the classification accuracy of entity identification may reach 95.85 percent, which is higher than the TBR-NER model's accuracy without transfer learning. Furthermore, the loss value of the TBR-NER model under transfer learning is as low as 5866, which is considerably less than the loss value of the four NER models discussed in Section 3.2. The smaller the loss value, the less the gap

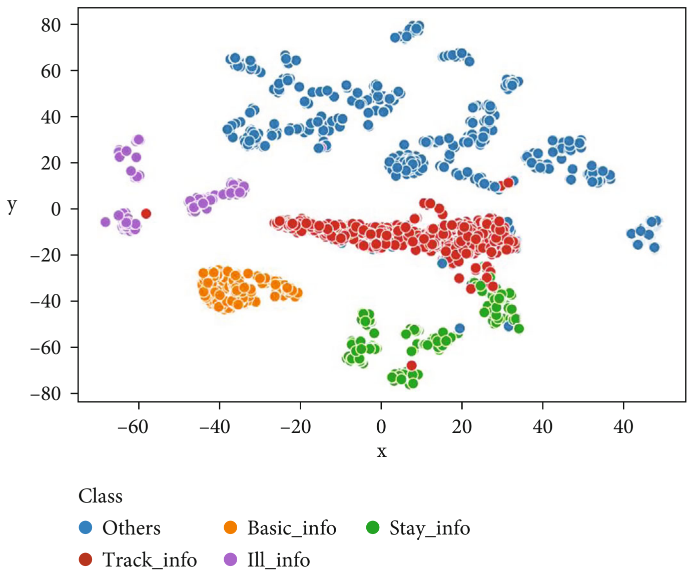


Figure 13: Visual display of 5 categories of topic t-SNE dimensionality reduction.

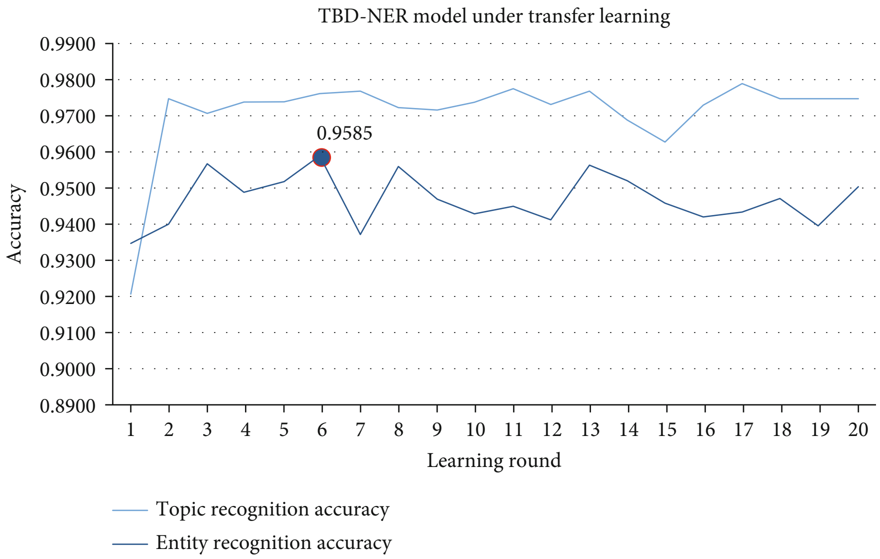


Figure 14: TBR-NER model under transfer learning.

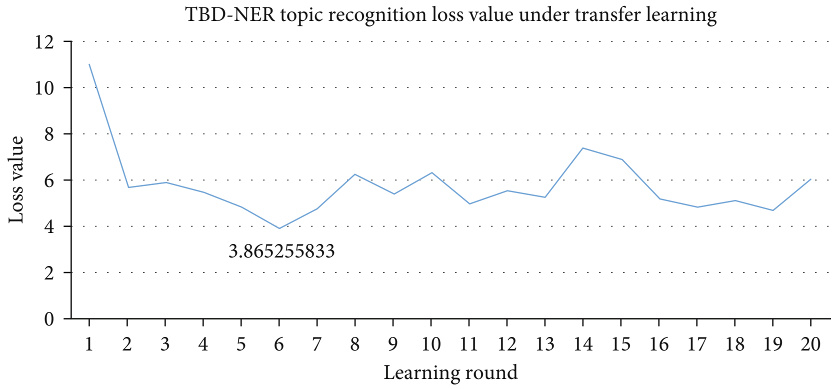


Figure 15: Loss value of TBR-NER topic recognition under transfer learning.

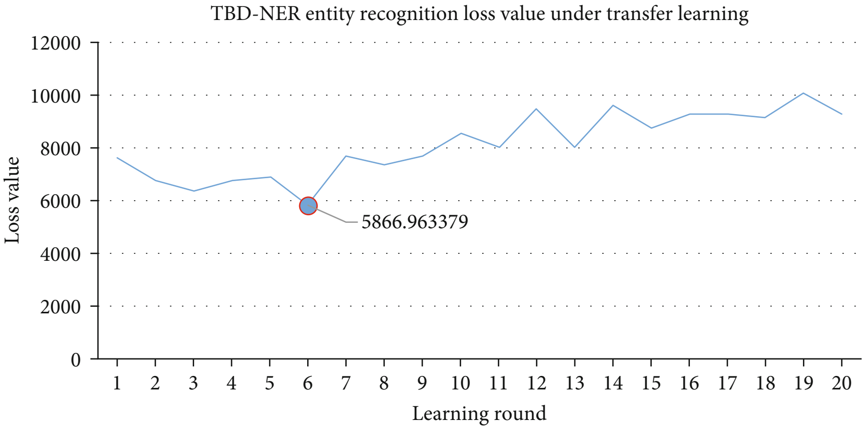


Figure 16: Loss value of TBR-NER entity recognition under transfer learning.

Table 3: Analysis and comparison of the TBR-NER model under transfer learning.

	TBR-NER model for topic recognition	TBR-NER model entity recognition
Accuracy rate	0.9791	0.9585
Loss value	3.8656	5866.963

Table 4: Comparison of TBR-NER under transfer learning and TBR-NER without transfer learning.

	Topic recognition of the TBR-NER model under transfer learning	Entity recognition of the TBR-NER model under transfer learning	Topic recognition of the TBR-NER model without transfer learning	Entity recognition of the TBR-NER model without transfer learning
Accuracy rate	0.9791	0.9585	0.9709	0.9160
Loss value	3.8656	5866.963	6.0467	11850.74

Table 5: Comparison of entity recognition between the TBR-NER model and other four NER models.

	BERT + BiLSTM + CRF	BERT + CRF	BiLSTM + CRF	LSTM + CRF	TBR-NER
Accuracy rate	0.4904	0.4904	0.7132	0.7098	0.9585
Loss value	73349.1 9	73936.48	28270.13	29563.14	5866.963

between the model's forecast and reality and the more convergent the model. Furthermore, a comparison of Figures 10 and 14 reveals that the accuracy of entity identification and topic recognition in the TBR-NER model impacts each other, an intuitive representation of the TBR-NER model's joint learning.

The analysis of topic recognition and entity recognition of the TBR-NER model under transfer learning is shown in Table 3. The comparison results of topic recognition and entity recognition between TBR-NER under transfer learning and TBR-NER without transfer learning are shown in Table 4. The comparison results of entity recognition between the TBR-NER model under transfer learning and the other four NER models are shown in Table 5.

Table 6: The prediction results were analyzed and compared.

Assessment method	Exact number	Total number	Accuracy
Word (unit)	16002	17644	0.91
Entity word (unit)	919	1052	0.87

4.4. Generalization of the TBR-NER Model. To verify the generalization of the TBR-NER model, this article uses 89 data from February 18 to March 1, 2021, in Hebei province's epidemic notification information as an independent test set. The word-to-word entity recognition accuracy is used as the evaluation index. Table 6 displays the accuracy rate. Table 6 shows that the prediction accuracy of outcomes with words as labels may reach over 90%, demonstrating that the proposed TBR-NER model has strong generalization.

5. Conclusions

This paper proposes a novel text information extraction method for COVID-19 outbreak notification based on joint learning of topic recognition and named entity recognition. The model can simultaneously complete the task of subject recognition and entity recognition, realize the subject self-labeling process based on the labeling results, and effectively improve the accuracy of the classification of epidemic notification information. This method employs rule matching and conditional random fields for word segmentation based on the regularization characteristics of epidemic information description. It uses data enhancement and transfers learning techniques to solve the model's generalization ability in the absence of samples and improve the analytical power of text information. The experiment is put up against four different natural language processing systems. The findings reveal that the suggested method's accuracy on the test set has dramatically improved and the loss function has been significantly lowered. Furthermore, the topic recognition task and the entity recognition task have many mutual promotions in the TBR-NER model's learning process, demonstrating that learning topic recognition and entity recognition together in text information extraction complement each other. The final model generated in this article is validated using datasets from Jilin and Hebei provinces. The concept will be expanded in the future to include datasets from more regions, giving important information for epidemic prediction and prevention.

In a future work, we plan to explore more countries' new crown epidemic notification information and text label information in other languages. Other countries and regions can use this model to predict their text trajectory information. Younes et al. [29] used NLP-related technologies to transcribe Arabic and Latin languages. After exploration and research, we can try to transplant the text information model to Arabic and African languages in the future to contribute to global epidemic prevention. Updating and improving the model to improve predictions' accuracy will significantly help in epidemic prevention and control. In addition, the model can be inherited in the GUI through pyqt5 and combined with the real-time map software to provide user convenience and epidemic prevention. Users can use the model to extract actual location words and path trajectories into the JSON data format and dynamically display the epidemic trajectory route map through web pages in real time.

Data Availability

The experimental data of this paper comes from the website, and the specific website is as follows: epidemic situation notification website of Hebei province, China, http://wsjkw.hebei.gov.cn/gzdt/index_43.jhtml, and epidemic situation notification website of Jilin province, China, http://wsjkw.jl.gov.cn/xwzx/xwzx/index_41.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

XF and QY conceived the project, designed the experiments, and drafted the manuscript. RH, YR, ZH, and ZF collected the data and conducted the experiments. RH, YR, and XF proofread and polished the manuscript and organized this project.

Acknowledgments

This work is supported by the science and technology project of the Education Department of Jilin Province (JJKH20220245KJ and JJKH20220226SK), the Higher Education Scientific Research Projects of Jilin Province (JGJX2021D213 and JGJX2021D226), the National Natural Science Foundation of China joint fund project (U19A200496), and the Ministry of Education's Industry-University Cooperation and Collaborative Education Project (202101244021).

References

[1] J. Wu, J. Wang, S. Nicholas, E. Maitland, and Q. Fan, "Application of big data technology for COVID-19 prevention and control in China: lessons and recommendations," *Journal of Medical Internet Research*, vol. 22, no. 10, article e21980, 2020.

[2] C. A. D. Durai, A. Begum, J. Jebaseeli, and A. Sabahath, "COVID-19 pandemic, predictions and control in Saudi Arabia using SIR-F and age-structured SEIR model," *The Journal of Supercomputing*, vol. 78, no. 5, pp. 7341–7353, 2022.

[3] Q. Ye, J. Zhou, and H. Wu, "Using information technology to manage the COVID-19 pandemic: development of a technical framework based on practical experience in China," *JMIR Medical Informatics*, vol. 8, no. 6, article e19515, 2020.

[4] J. Li, S. Zhao, J. Yang et al., "WCP-RNN: a novel RNN-based approach for bio-NER in Chinese EMRs," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1450–1467, 2020.

[5] S. Bhaskar, S. Bradley, S. Sakhamuri et al., "designing futuristic telemedicine using artificial intelligence and robotics in the COVID-19 era," *Frontiers in Public Health*, vol. 8, article 556789, 2020.

[6] S. J. Alsunaidi, A. M. Almuhaideb, N. M. Ibrahim et al., "Applications of big data analytics to control COVID-19 pandemic," *Sensors*, vol. 21, no. 7, article 2282, 2021.

[7] M. Khayyat, K. Laabidi, N. Almalki, and M. Al-Zahrani, "Time series facebook prophet model and python for COVID-19 outbreak prediction," *Computers, Materials, & Continua*, vol. 67, no. 3, pp. 3781–3793, 2021.

[8] J. Song, H. Xie, B. Gao, Y. Zhong, C. Gu, and K.-S. Choi, "Maximum likelihood-based extended Kalman filter for COVID-19 prediction," *Chaos, Solitons & Fractals*, vol. 146, article 110922, 2021.

[9] H. M. Balaha, E. M. el-Gendy, and M. M. Saafan, "CovH2SD: a COVID-19 detection approach based on Harris hawks optimization and stacked deep learning," *Expert Systems with Applications*, vol. 186, article 115805, 2021.

[10] M. Wieczorek, J. Siłka, and M. Woźniak, "Neural network powered COVID-19 spread forecasting model," *Chaos, Solitons & Fractals*, vol. 140, article 110203, 2020.

[11] K. Kozioł, R. Stanisławski, and G. Bialic, "Fractional-order SIR epidemic model for transmission prediction of COVID-19 disease," *Applied Sciences*, vol. 10, no. 23, p. 8316, 2020.

[12] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Gesture recognition using latent-dynamic based conditional random fields and scalar features," *Journal of Physics: Conference Series*, vol. 812, article 012113, 2017.

[13] A. Allam and M. Krauthammer, “PySeqLab: an open source Python package for sequence labeling and segmentation,” *Bioinformatics*, vol. 33, no. 21, pp. 3497–3499, 2017.

[14] F. Zhuang, Z. Qi, K. Duan et al., “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[15] A. Khamparia, D. Gupta, V. H. C. de Albuquerque, A. K. Sangaiah, and R. H. Jhaveri, “Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning,” *The Journal of Supercomputing*, vol. 76, no. 11, pp. 8590–8608, 2020.

[16] M. Zhang, G. Geng, S. Zeng, and H. Jia, “Knowledge graph completion for the Chinese text of cultural relics based on bidirectional encoder representations from transformers with entity-type information,” *Entropy*, vol. 22, no. 10, p. 1168, 2020.

[17] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors, “Using rule-based natural language processing to improve disease normalization in biomedical text,” *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 876–881, 2013.

[18] T. Kang, A. Perotte, Y. Tang, C. Ta, and C. Weng, “UMLS-based data augmentation for natural language processing of clinical research literature,” *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 812–823, 2021.

[19] C. Mi, S. Zhu, and R. Nie, “Improving loanword identification in low-resource language with data augmentation and multiple feature fusion,” *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 9975078, 9 pages, 2021.

[20] J. Wei and K. Zou, “EDA: easy data augmentation techniques for boosting performance on text classification tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382–6388, Hong Kong, China, 2019.

[21] X. Xing, R. Xie, and W. Zhong, “Model-based sparse coding beyond Gaussian independent model,” *Computational Statistics & Data Analysis*, vol. 166, article 107336, 2022.

[22] M.-F. Tsai and H.-J. Tseng, “Enhancing the identification accuracy of deep learning object detection using natural language processing,” *The Journal of Supercomputing*, vol. 77, no. 7, pp. 6676–6691, 2021.

[23] T. H. V. Phan and P. Do, “BERT+vnKG: using deep learning and knowledge graph to improve Vietnamese question answering system,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[24] A. H. A. Rahnama, M. Toloo, and N. J. Zaidenberg, “An LP-based hyperparameter optimization model for language modeling,” *The Journal of Supercomputing*, vol. 74, no. 5, pp. 2151–2160, 2018.

[25] Y. Song, S. Tian, and L. Yu, “A method for identifying local drug names in Xinjiang based on BERT-BiLSTM-CRF,” *Automatic Control and Computer Sciences*, vol. 54, no. 3, pp. 179–190, 2020.

[26] G. Yu, Y. Yang, X. Wang et al., “Adversarial active learning for the identification of medical concepts and annotation inconsistency,” *Journal of Biomedical Informatics*, vol. 108, no. 2, article 103481, 2020.

[27] Y. Chen, C. Zhou, T. Li et al., “Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training,” *Journal of Biomedical Informatics*, vol. 96, article 103252, 2019.

[28] C. Lee, “LSTM-CRF models for named entity recognition,” *IEICE Transactions on Information and Systems*, vol. 100, no. 4, pp. 882–887, 2017.

[29] J. Younes, E. Souissi, H. Achour, and A. Ferchichi, “Language resources for Maghrebi Arabic dialects’ NLP: a survey,” *Language Resources and Evaluation*, vol. 54, no. 4, pp. 1079–1142, 2020.