scientific reports

OPEN

Information extraction from green channel textual records on expressways using hybrid deep learning

Jiaona Chen[1]✉, Jing Zhang[1], Weijun Tao[1], Yinli Jin[2] & Heng Fan[1]✉

The expressway green channel is an essential transportation policy for moving fresh agricultural products in China. In order to extract knowledge from various records, this study presents a cutting-edge approach to extract information from textual records of failure cases in the vertical field of expressway green channel. We proposed a hybrid approach based on BIO labeling, pre-trained model, deep learning and CRF to build a named entity recognition (NER) model with the optimal prediction performance. Eight entities are designed and proposed in the NER processing for the expressway green channel. three typical pre-trained natural language processing models are utilized and compared to recognize entities and obtain feature vectors, including bidirectional encoder representations from transformer (BERT), ALBERT, and RoBERTa. An ablation experiment is performed to analyze the influence of each factor on the proposed models. Used the survey data from the expressway green channel management system in Shaanxi Province of China, the experimental results show that the precision, recall, and F1-score of the RoBERTa-BiGRU-CRF model are 93.04%, 92.99%, and 92.99%, respectively. As the results, it is discovered that the text features extracted from pre-training substantially enhance the prediction accuracy of deep learning algorithms. Surprisingly, the RoBERTa model is highly effective in the task for the expressway green channel NER. This study provides a timely and necessary knowledge extraction on the Expressway Green Channel in terms of textual data, offering a systematical explanation of failure cases and valuable insights for future research.

Keywords Expressway green channel, Named entity recognition, BIO labeling, Pre-trained model, Deep learning

The Green Channel for fresh agricultural products on the expressway is an essential policy in China that involves all toll stations in the country and millions of toll-free vehicles. The Green Channel was first implemented in 1995. Since 2010, it has been extended to all toll roads in China[1]. The toll for freight vehicles on expressways in China is calculated by total weight of the vehicle and goods. A special lane, called the Green Channel, is set up for vehicles at the entrance and exit of each toll station, which loads fresh agricultural products. The Green Channel vehicles are designed to carry fresh, low-cost, and short-life agricultural products on expressways. The catalog of fresh agricultural products includes fresh vegetables, fruits, aquatic products, live livestock, poultry, fresh meat, eggs, and milk. Here, we briefly introduce the inspection process of the Green Channel, which is defined to check the truck belongs to the Green Channel vehicle or not. Green Channel vehicles are the same as any other regular trucks when entering the expressway. The trucks will be inspected at the toll station exit of the Green Channel. If a vehicle passes the inspection process when exiting the toll station, it is defined as a Green Channel vehicle. The toll for a Green Channel vehicle can be waived for reducing the cost of fresh agricultural products on transportation. The drivers will be exempt from the toll if they pass the inspection. In the event of inspection failure, the drivers will be responsible for paying the toll and their names will be included in a grey list. Weighing the vehicle and goods, checking certificates, and taking photos of the vehicle and goods are the necessary inspection procedures. If necessary, the drivers will be required to assist with unloading the truck to check whether the agricultural products meet the free standard. The main aim of the Green Channel is to reduce tolls for those trucks transporting fresh agricultural products, allowing these agricultural products to go through the toll station without unloading and facilitating their transportation on time at low costs. This measure is

[1]Xi'an Shiyou University School of Electronic Engineering, Xi'an 710065, China. [2]Chang'an University School of Electronic and Control Engineering, Xi'an 710065, China. ✉email: chenjn@xsyu.edu.cn; fan_h@xsyu.edu.cn

essential for keeping produce fresh and reducing transport costs. This Green Channel policy also contributes to the stability of local prices, making it appealing for producers and consumers. The "vegetable basket" price has been regulated for over 20 years, thanks partly to the Green Channel implementation.

However, this system has strict criteria that must be met to function effectively. These criteria are outlined in three main aspects[2]. First, only light, fresh agricultural products are included in the listed catalog. Second, mixed loading with non-fresh agricultural products is permitted. Moreover, the loaded fresh agricultural products should account for more than 80% of the approved vehicle's authorized load mass or carriage volume. Finally, the green channel policy does not apply to vehicles that exceed the limit, are illegally overloaded, refuse to pass designated lanes, or object to inspection. Similar restrictions are applied to vehicles on the grey list for a specific period.

It is essential for extracting experiential knowledge from historical records to achieve economic and efficiency objectives. Some problems have arisen regarding the Expressway Green Channel implementation. The problems are long queues for transporting green products when passing the toll station inspection, especially during heavy traffic flow. As a part of the inspection procedures, checking Green Channel vehicles at the toll station departure is a routine and complex task for the workers. Exploit the convenience in this policy, numerous unlawful practices of counterfeiting Green Channel goods have been used to evade taxes. Adequate inspection time will be beneficial to find the cheating behaviors, but it will also affect more frequent congestion at the toll station. A balance between economy and efficiency will be established using the experiential knowledge extracting.

The toll waivers, especially for trucks, are substantial, and drivers closely monitor whether the exemption policy applies to them. Additionally, Green Channel is attracting considerable critical attention because its inventory of fresh agricultural products has not been updated frequently. Due to regional differences, the inspection outcomes for the same case may be different by different local toll road management departments. However, these results may be different in different regions, inconsistent understanding of Green Channel can cause some negative social comments, and ignoring these comments could damage the significance of this policy. Consequently, information extraction from green channel records would be highly useful for guiding transportation agencies and drivers in the toll station inspection. Firstly, lots of agricultural products that are not easily distinguishable can be learned by toll station staffs and drivers. Secondly, some special and rare cases would be shown in details. Finally, it helps to eliminate the deviation of policy understanding between toll station staffs and drivers by case study.

An electronic record of each truck through the Green Channel will be saved whether it meets the standards or not. Previously, toll station staff manually recorded information to show evidence of fee waivers. As the lack of detailed failure case description, information extraction from Green Channel records on expressways has been a difficult problem for many years. In recent years, Green Channel inspection has become increasingly digitalized. Truckers and toll station employees use mobile phones to make electronic recordings because knowledge and experience are accumulated thanks to these electronic records. Notably, natural language describes the reason for inspection disqualification. For failure cases, the reason for the failure or other noteworthy information will be recorded in natural language. Multiple failed records will affect the credit of drivers.

With the recent application of artificial intelligence to the transportation field, it is expected to handle unstructured data for the expressway green channel management system. The unstructured text conveys descriptions of causes and manifestations on failure cases. As the electronic textual records has been adopted, it provide the possibility for extracting experience knowledge. However, few studies have focused on these issues despite more drivers and managers being concerned about these problems. Moreover, characteristic analysis and information extraction of failure cases are helpful to implement the policy. Then it is an urgent task to information extraction for experience knowledge integration from green channel textual records on expressways.

This study proposes extracting experience and knowledge from unstructured text data to analyze the failure cases of the Green Channel on the expressway toll-free. Within the standard framework in the context of Chinese NER tasks, our study provides guidance on best practices for Green Channel Named Entity Recognition. Furthermore, it derives additional insights from practical applications of Expressway Green Channel Management data. Moreover, our findings can help to improve the knowledge of drivers and toll station staff on toll-free rules application and greatly enhance their efficiency in applying toll-free rules.

Related work

Information extraction is a critical area in natural language processing. However, scholarly articles discussing specifically Green Channel information extraction have not been easy to find. Instead of information extraction on the Expressway Green Channel, there are a higher number of scholarly articles focused on named entity recognition in other areas. In addition, named entity recognition (NER), the focus of previous studies on information extraction, is the task of finding the position of proper names in a sentence and assigning it to the correct category.

In previous years, the primary approach for NER was based on rules and dictionary solutions. Lee et al.[3] demonstrated the effectiveness of conditional random fields (CRFs) for performing named entity recognition (NER) on clinical and general datasets. Yi et al.[4] proposed a novel approach to NER in security text data from public security webs, utilizing regular expressions, known-entity dictionaries, CRFs, and four feature templates. Furrer et al.[5] developed OGER++, a text-mining tool that achieved high accuracy in recognizing biomedical entities on the CRAFT corpus, with 71.4% and 56.7% F1 for named entity recognition and concept recognition, respectively .

Recently, the neural network has significantly dominated this field with the rise of machine learning. In Han 's research[6], the proposed multifeature adaptive fusion Chinese named entity recognition (MAF-CNER) model utilized bidirectional long short-term memory (BiLSTM) neural network to achieve superior performance, with F1 values of 97.01% and 96.78% on Microsoft Research Asia (MSRA) and "China People's Daily" dataset from January to June 1998, respectively. The bidirectional long short-term memory–conditional random field (BiLSTM-CRF) approach, as the main structure, is widely used in NER fields and has achieved remarkable results. Shi et al.[7] proposed a unified framework by integrating BiLSTM-CRF and BiLSTM, which outperform the other methods on entity recognition tasks and relation extraction tasks. Chen et al.[8] combined the BiLSTM-CRF with an improvement of sequence annotation rules, with the accuracy rate of 83.46%, the recall rate of 81.12%, and the F1 value of 0.8227. Qi et al.[9] suggested an improved training algorithm combining Iterated Dilated Convolutional Neural Networks with BiLSTM for the low-resource Chinese medical entity recognition task. Kang et al.[10] integrated character boundary information based on BiLSTM-CRF to significantly enhance the attention network and improve the model's recognition of entity boundaries. The experiments results show that the proposed method improves the F1 value by 2% on average in the field with low-resource based on the ACL paper data set, China referee network desensitization data set and ccks2018 data set. Kim et al.[11] developed a neural biomedical named entity recognition and multi-type normalization tool called BERN for biomedical text mining tasks, such as new named entity discovery, information retrieval, question answering, and relation extraction.

As Gong et al. reported[12], the deep learning method, pre-training model has become a hot research topic in many NPL fields, such as Encoder Representations from Transformer (BERT) and Robustly Optimized BERT (RoBERTa). Pre-training models were introduced into entity recognition for better performance. Gao et al.[13] introduced the BERT model to relation extraction and entity recognition, and obtained excellent results by comparing with most existing models on the New York Times (NYT) and WebNLG datasets. Chen et al.[14] improved the fine-grained geological NER using BERT-Flat-Lattice Transformer, achieving better performance than BERT-CRF as the F1-score of 92.05%. Fang et al.[15] suggested a multi-task learning network MTL-BERT to improve the model's feature extraction abilities.

Some scholars have tried to improve the NER method on the basis of attention mechanisms. He et al.[16] propose a novel approach for enhancing the performance of named entity recognition by integrating knowledge graph embedding and self-attention mechanism. The experimental results demonstrate its effectiveness on marine corpus as well as other publicly available datasets. Wang et al.[17] conducted a cross-lingual entity recognition method that leverages attention mechanism and adversarial training, achieving the optimal value of 53.22% on WeiboNER dataset and People-Daily2004 dataset.

However, these existing models face challenges in offering good performance owing to the complexity and specificity of Chinese text. Other strategies have also been tested to extract textual features in Chinese NER. He et al.[18] performed an innovative association rule mining technique based on named entity recognition and text classification (ARMTNER), achieving an impressive F1-score of 97.3% on a public Chinese Named Entity Recognition dataset. Sun et al.[19] obtained a natural hazard-NER based on the XLNet-BiLSTM-CRF model. Their proposed approach outperformed alternative methods in identifying research hotspots of natural hazards papers over a decade-long period, achieved impressive precision (92.80%), recall rate (91.74%), and F1-score (92.27%).

Furthermore, the nested named entities, where one entity is contained within another, have received significant attention from scholars[20]. Li et al.[21] presented an innovative approach for named entity recognition (NER) utilizing global pointer mechanism combined with adversarial training techniques. Through extensive experiments conducted on widely-used public datasets such as OntoNotes5, MSRA dataset al.ong with Resume dataset and Weibo dataset, they observed significant improvements in terms of F1 scores when compared to existing mainstream models. Additionally, Li et al.[22] focused specifically on a head-to-tail linker for nested NER. Their findings demonstrated the effectiveness achieving F1-scores of 80.5%, 79.3%, and 76.4% within given text data sets including ACE2004, ACE2005, and GENIA. Gao et al.[23] verifies that the proposed method has better performance on the nested entity recognition task, which improve the F1 value by up to 7.05%, 12.63%, and 14.68% on the GENIA, ACE2004, and ACE2005 nested datasets. Zhang et al.[24] proposes a novel approach using Large Language Models (LLMs) to systematically extract and analyze the event chain within TCE.

Compared with the above NER researches, the study on the Expressway Green Channel management data is inadequate. It is necessary to extract information from practical implementation experience for the complexity and diversity in real scenarios. In our previous study[25], the cause mechanism of unqualified toll-free vehicles for the Green Channel has been analyzed using text mining based on dictionary solutions. As the actual cases are complex and varied, the limited dictionary of the Green Channel cannot cover all situations. As a result, it is insufficient to support subsequent natural language processing tasks.

In order to improve the capability of the model, this study proposes a named entity recognition model for the complexity and variability of text descriptions. It aims to explore the policy in detail to give drivers information for them to figure out if they qualify for an exemption or how to load to be eligible for toll-free. This innovative study explores the causes of unqualified vehicle analysis and will contribute to a deeper understanding of unqualified vehicles. In particular, failure cases will assist drivers in enjoying preferential policies in terms of avoiding similar situations.

In Expressway Green Channel management data, we used natural language descriptions of failure cases as the data source. A mixed-method approach based on a deep learning model was used to extract NER information. First, we used the pre-training model to vectorize irregular text information. Second, we analyzed a combination of deep learning approaches to reduce error. Third, we opted for the CRF approach to output optimization results. Afterward, we evaluated the model's performance, conducting elimination tests to verify the proposed model's superiority. Finally, we established the proposed NER model for the Expressway Green Channel.

This study offers the following contributions:

(1) In a relatively new study in the field of the Expressway Green Channel, we thoroughly investigated the failure cases from the Expressway Green Channel management departments to acquire a processing technique for unstructured text.

(2) This study adopts a hybrid framework based on RoBERTa-BiGRU-CRF to extract information for Expressway Green Channel NER task. As the small field of NLP research, the guidelines for best practices are valuable in Green Channel applications.
(3) Evaluation results show that the proposed method achieves better results on the comprehensive F1-score than existing mainstream pre-trained and deep-learning models.

Methodology

Named entity recognition framework

The expressway green channel dataset consists of electronic records from all vehicles through the Green Channel at the exit of the toll station, including vehicle information, driver information, inspection photos, and cargo information. In particular, the unqualified vehicle will be recorded as the reason for the failure in natural language. As a result, these natural language texts are the data sources for this study. This dataset is collected and managed by governmental departments for the management of toll on expressways. As recorded dataset in the real operation, we acquired it after declassification from management departments for research purposes.

Based on text information, the hybrid model of expressway green channel NER contains four modules, including a text annotation module, a pre-trained language module, a deep learning module, and a CRF module. Firstly, the Begin-Intermediate-Other (BIO) labeling was used to mark the text of the Expressway Green Channel dataset in the text annotation module. Secondly, feature vectors incorporating word, sentence, and position vectors are generated using a bidirectional Transformer coding structure in the pretraining language module to obtain the vectorized representation of text sequences. Thirdly, the pre-trained language model generates the text vector, which is input into the deep learning model to extract features. Finally, using a deep learning model, we used the CRF to encode and restrict the sequence annotation output. Accordingly, we obtained the sequence annotation with the highest probability as the final output.

The mechanism for using several models as a hybrid approaches is depicted as an ensemble architecture in Fig. 1. It is shown that the hybrid named entity recognition framework is achieved based on BIO labeling, pre-trained model, deep learning and CRF. The methodology of our study is distinctive and innovative on extracting the knowledge of Green Channel on expressways from textual data, as well as offering empirical conclusions.

Text annotation based on begin-intermediate-other

We used the BIO model for text annotation of unstructured natural language descriptions, where B stands for the entity's beginning, I for its middle, and O for its non-entity. The text annotation based on BIO model comprises three distinct processes: (1) word segmentation; (2) entity definition and coding; (3) BIO labeling. The BIO annotation process of textual data is the normal task in NER. The novel contribution of our methodology is found in the rational design of the entity for Expressway Green Channel.

Word segmentation

The text is processed by word segmentation processing before annotation. Word segmentation is a crucial step in Chinese NER, involving the division of continuous Chinese text into independent words or terms. In practical applications, widely utilized tools such as jieba are employed for Chinese word segmentation. To enhance the proposed model's comprehension of fundamental elements and semantics in the text, we have adopted the precise mode of jieba following the definition of a specialized dictionary for Green Channel text.

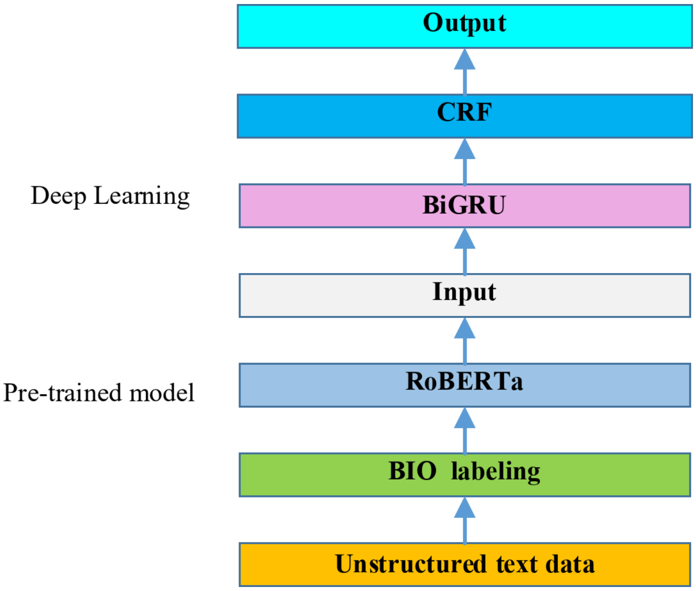

Fig. 1. The architecture of named entity recognition for expressway Green Channel.

Entity	Code	Annotation of the beginning for each entity	Annotation of the middle for each entity	Textual examples
Cheating	Entity 1	B-Entity 1	I-Entity 1	假冒, 伪造, 涂改
Mixed	Entity 2	B-Entity 2	I-Entity 2	混装超过20%
Out-listed	Entity 3	B-Entity 3	I-Entity 3	快递, 榴莲, 菠萝蜜
Not fresh	Entity 4	B-Entity 4	I-Entity 4	非鲜活, 不新鲜
Frozen	Entity 5	B-Entity 5	I-Entity 5	冷冻, 冻硬
Spoilage	Entity 6	B-Entity 6	I-Entity 6	变质, 烂果, 发霉
Over-loading	Entity 7	B-Entity 7	I-Entity 7	外廓尺寸超限, 超载
Deep-processing	Entity 8	B-Entity 8	I-Entity 8	深加工, 卤制, 腌制
Non-entity	O	O		该, 拉, 运

Table 1. The entity definition and annotation.

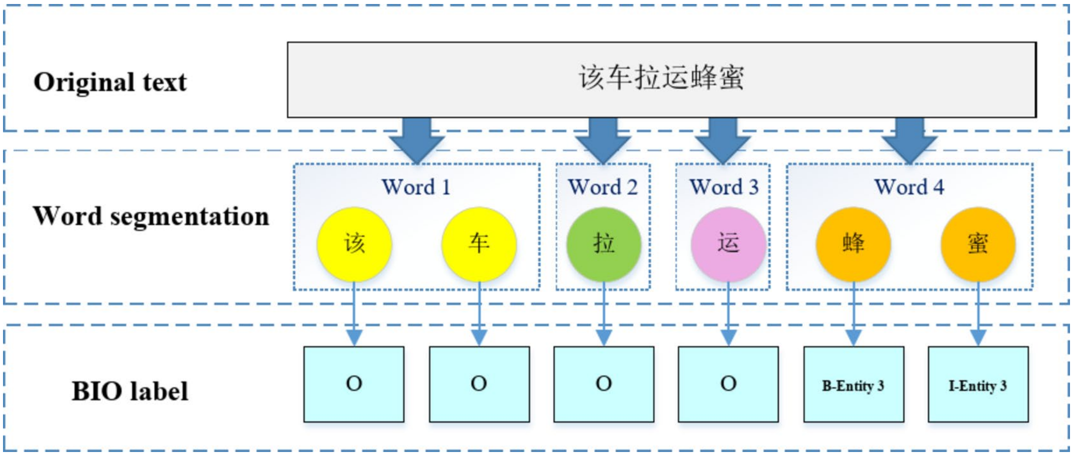

Fig. 2. The annotation processing of the BIO label.

Entity definition and coding

We developed and implemented eight entities in alignment with the research objectives. The code and annotation for Expressway Green Channel entities are shown in Table 1. Some textual examples are provided for understanding friendly in Table 1. Thus, the BIO label is integrated with the related named entity.

BIO labeling

As shown in Table 1 and 17 labels for BIO annotation were obtained. Each BIO label is set as an output category in the prediction model. Each word is labeled by the BIO annotation process. Figure 2 displayed the annotation effect by BIO labeling. Firstly, the sentence is divided into four words after word segmentation. Secondly, the fourth word is annotated as "B-Entity 3" and "I-Entity 3". the label of "B-Entity 3" indicates that the annotated sequence is the beginning for the "Out-listed" entity in further prediction model.

BERT is a deeply bidirectional language model that can be used to pre-train text features using the Masked Language Model and Next Sentence Prediction. The BERT model involves stacking multiple transformer encoders for feature extraction. In the BERT paradigm, the feature extraction is accomplished by stacking multiple Transformer encoders. First, [CLS] is inserted at the beginning of the text, and [SEP] is inserted between sentences. Embedding codes comprise the feature vector, word vector, sentence vector, and position vector. Mask words were chosen randomly with a proportion of 15% input into the BERT model before the feature vector of the text sequence. In 80% of the samples, the mask words were replaced with mask labels [MASK]. For the remaining 10% of the samples, the terms from the model glossary were randomly chosen in place of the mask words. The final 10% of samples would have nothing replaced. However, the BERT model can accurately predict these masked words to extract semantic information from the text data by conducting self-supervised learning training on large-scale corpus data.

As a result, the output vector corresponding to the [CLS] tag is denoted as C. If sentence B is the next sentence after sentence *A*, then *C* has the value of 1. Otherwise, *C* has a value of 0. TMask is the predicted mask word in which vector E={E_1,E_2…, E_N} is input into the BERT model after superposition and mask operation.

Each Transformer encoder consists of a self-attention layer and a feedforward neural network layer. Moreover, self-attention is a central mechanism in Transformer, which linearly transforms the input word vector E into three vector sequences: Q, K, and V. Then, the softmax function is used to obtain the feature function $Attention(Q, K, V)$based on the following computation:

$$Q = \mathrm{W}_q E \tag{1}$$

$$K = \mathrm{W}_k E \tag{2}$$

$$V = \mathrm{W}_v E \tag{3}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \tag{4}$$

where d_k is the dimension of word vector E, and W_q, W_k, and W_v are weight matrices obtained from training.

After pre-training by the BERT model, having the same dimension as the input word vector E, the output vector T={T_1,T_2… T_N} is generated by repeating the Transformer encoder several times in the same way,

ALBERT and RoBERTa have robustly optimized BERT that uses sentence-order prediction (SOP) tasks instead of NSP. ALBERT also uses Factorized Embedding Parameterization and Cross-layer Parameter Sharing.

As a result, ALBERT is more robust than BERT because its memory is more affordable, uses lower computing power, and is a lightweight version of BERT. Thus, ALBERT is more advantageous than BERT in training speed. RoBERTa has a grander scale in the model parameters, more batch size, and training data than BERT because of its alternative improvement technique. However, the training does not include NSP tasks, and when a sequence is input into the model, a new mask is generated each time. To learn different language representations, the RoBERTa model can gradually deal with new dynamic masks, especially when the mask position is continuously changing.

Thus, no pre-trained model can successfully offer a decent model size, have advanced small sample capability, and provide fine-tuning capability based on the outcome of the analysis. Numerous studies revealed that each pre-trained model has unique benefits and shortcomings. Thus, each model must be verified during application. The following section determines the effectiveness of the pretraining model.

BiGRU-CRF model

Textual records in the Green Channel often encompass a diverse range of agricultural product names or abbreviations. Distinct personnel contribute to distinct textual characteristics associated with the same product. The existing structure of the BiGRU-CRF model is taken in our study. The BiGRU model is employed to capture contextual information for recognition labeling. Considering the interdependence between labels, the CRF model is utilized to ensure the global optimality of predicted label sequences. By synergistically integrating the advantages of BiGRU and CRF, the BiGRU-CRF model effectively captures contextual information in Green Channel textual records during named entity recognition tasks, thereby enhancing both accuracy and efficiency of recognition. For superior performance, the comprehensive experimental analysis for our specific task is essential. The proposed BiGRU-CRF model is gained from empirical analysis by fine-tuning some or all of the parameters.

After pretraining, the vector T is processed using the BiGRU model to extract features. Then, the CRF model constrains the BiGRU model output to obtain accurate prediction results. Figure 3 depicts the proposed prediction model based on the BiGRU-CRF architecture, indicating that the BiGRU-CRF framework consists of four layers: an input layer with a word vector after pretraining, a BiGRU layer that determines the predicted probability for each entity type, a CRF layer with constraint condition, and an output layer.

As illustrated in Fig. 4, the BiGRU model consists of forward GRUs and backward GRUs to acquire the contextual features in both directions. The first forward GRU processes the forward information of the sequence, while the first reverse GRU processes the reverse information of the sequence. Each GRU's output in both stages is combined to present the final and accurate results. Thus, the Softmax function presents the probability value p_{ij} for each word T_i corresponding to each predetermined label j. In the case of overfitting, the Dropout layer is adopted to the BiGRU network.

The GRU model is a variant of the LSTM model with a simpler network topology and fewer parameters than the LSTM model. A GRU model's structure consists of a hidden state H_{t-1} at time $t-1$, an input X_t and a hidden state H_t at time t, along the sigmoid σ and tanh activation functions. An update gate and a reset gate in the GRU model have values of U_t and R_t at time t, respectively, specified as follows:

$$R_t = \sigma\left(w_r \cdot [H_{t-1}, X_t] + b_r\right) \tag{5}$$

$$U_t = \sigma\left(w_u \cdot [H_{t-1}, X_t] + b_u\right) \tag{6}$$

Here, the terms $[H_{t-1}, X_t]$ represent the concatenation of vectors H_{t-1} and X_t, In addition, w_r and w_u are weight matrices, and b_r and b_u are bias values. Then, the process yields the candidate set of the current state and the value of H_t as follows:

$$\tilde{H}_t = \tanh\left(w_h \cdot [R_t * H_{t-1}] + b_h\right), \tag{7}$$

$$H_t{=}U_t * H_{t-1} + (1 - U_t)\,\tilde{H}_t, \tag{8}$$

where ω_h and b_h are the corresponding weight matrix and bias value, respectively. The weight matrices and bias values are determined during the training process.

Accordingly, the output $\vec{H}_t$ of a forward GRU and the output of a reverse GRU are presented as follows:

$$\vec{H}_t = GRU\left(X_t, \vec{H}_{t-1}\right) \tag{9}$$

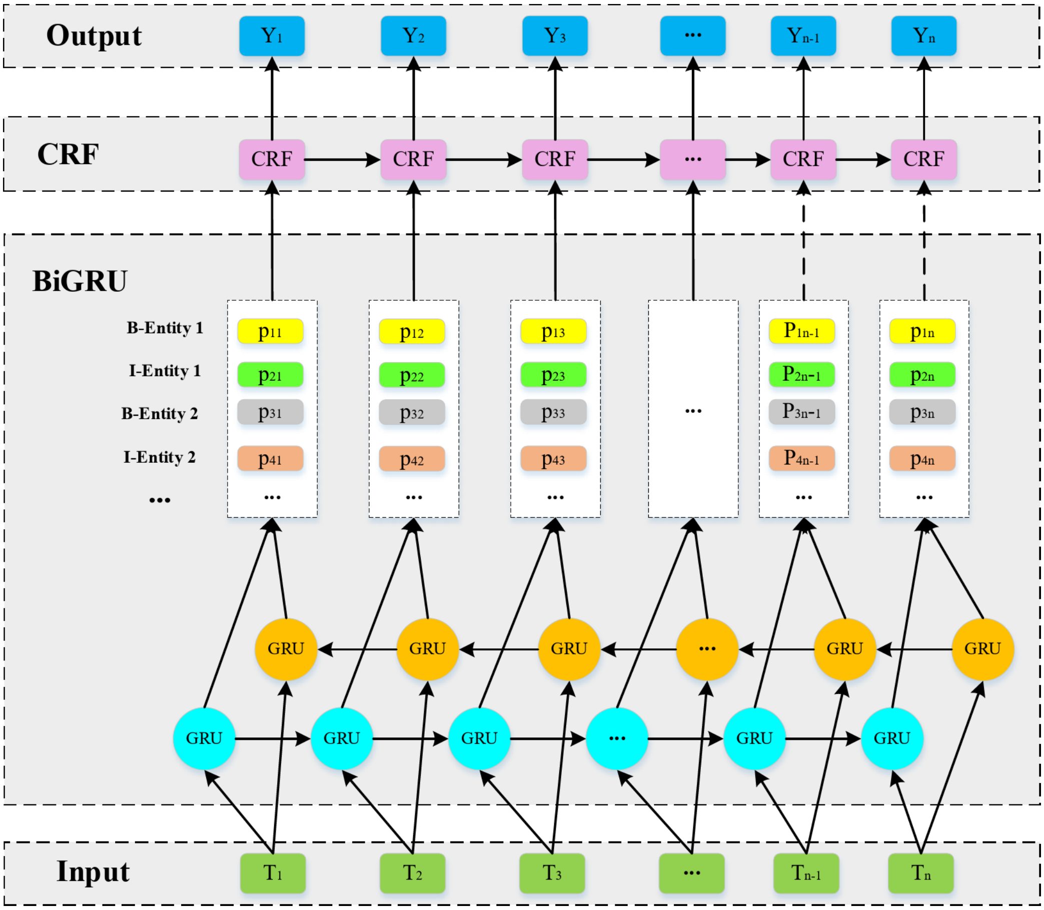

Fig. 3. The proposed prediction model based on BiGRU-CRF architecture.

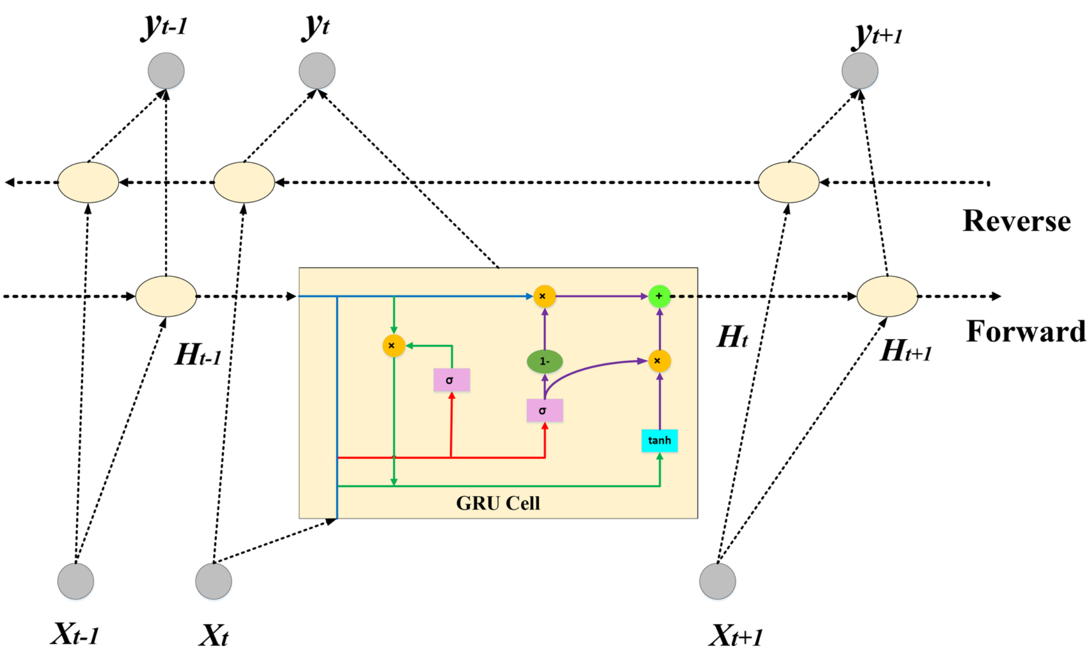

Fig. 4. The structure of GRU model.

$$\overleftarrow{H}_t = GRU\left(X_t, \overleftarrow{H}_{t-1}\right) \tag{10}$$

$$H_t = \left[\vec{H}, \overleftarrow{H}\right] \tag{11}$$

Here, *GRU*(·) represents the output of (4), and the final output is a combination of $\vec{H}_t$ and $\overleftarrow{H}_t$. The output vector $H = [\ H_1, H_2, \ldots, H_t]$ of all GRU sets in the BiGRU model obtained from (7) serves as the CNN's inputs, as illustrated in Fig. 2. The final output *Y* obtained after applying the softmax activation function *soft*(·) is given as follows:

$$Y = soft\left(W \cdot H + b\right), \tag{12}$$

where *W* is the weight matrix, and *b* is the bias.

Results

Data collection

Shaanxi Expressway, referred to as the expressways of Shaanxi Province, is an important part of China's national expressway network, as well as a transportation hub for moving fresh agricultural products from Northwest to Southwest China. The survey data comprised Green Channel expressway recordings gathered in Shaanxi Province between January 2020 and June 2022. The discrimination results, exit toll station, vehicle weight, and other attribute fields are reported during this data collection. Furthermore, the original data describes information in predefined and natural language fields linked to justifications for failure cases. As the self-constructed dataset by the Shaanxi Province Expressway Toll Center, it is a unique and abundant dataset in Expressway Green Channel fields. The dataset is also used in our previous study[25]. As discussed previously, we aim to describe the natural language for failure cases in original data to demonstrate the practical significance of data mining on these text data. After declassified and desensitized from the expressway management departments, we acquired the textual dataset for further study.

In the original dataset of Expressway Green Channel, 26,099 samples are recorded in the natural language descriptions. After eliminating empty values from the original dataset, 14,844 text records were deemed valid samples. As a result, the experimental data was generated as sample dataset on the Green Channel failure cases for our analysis.

The statistical description of samples was analyzed as shown in Table 2. Both the length of the text and the frequency of words are calculated in the statistical analysis. Text length refers to the length of the textual sequences for each sample. Word frequency refers to the number of times a specific word appears within our corpus.

According to the entities definition of Expressway Green Channel as shown in Table 1, the proportions of each entity are displayed in Fig. 5 for the sample dataset following text annotation based on BIO. It is shown that the proportion of the named entities is not evenly distributed in the sample dataset. In order to train the proposed model, 80% of the data was randomly selected as the training dataset, and the remaining 20% served as the testing dataset. Each experiment used identical training and testing dataset. The distribution of each entity for testing dataset is shown in Fig. 5. We can observe that the distribution of entities on the testing dataset is equally unbalanced as well as whole sample dataset. This shows that the random sampling method is reasonable and effective, as retained the unbalanced entity distribution.

Performance index

We evaluated the performances of the prediction models according to the precision P_j, recall R_j, and F_j metric values of categories *j*, calculated as follows:

$$P_j = \frac{TP_j}{TP_j + FP_j} \tag{13}$$

$$R_j = \frac{TP_j}{TP_j + FN_j} \tag{14}$$

$$F_j = \frac{2P_jR_j}{P_j + R_j} \tag{15}$$

	Max	Min	Mean	Standard deviation	Median	Kurtosis	Skewness	Coefficient variation
Text length	172	1	10.302	9.486	7	16.253	3.066	0.921
Word frequency	6141	1	11.164	141.970	1	1418.117	35.958	12.717

Table 2. The statistical description of samples.

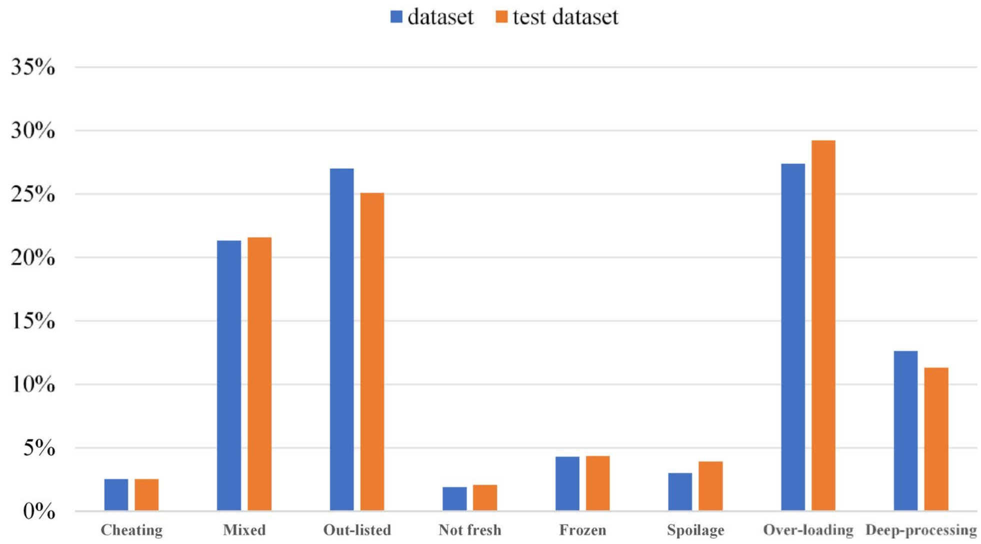

Fig. 5. The proportions for each entity in the dataset and the test dataset.

Parameter	Value
Vector dimension of BERT	768
Maximum sentence length	172
Hidden layer dimension of GRU or LSTM	768
Optimizer	Adam
Learning rate	5×10^{-5}
Batch size	16
Maximum iteration rounds	3

Table 3. The suggested super parameters of pre-trained models.

Here, TP_j is the true positives number, FP_j is the false positives, and FN_j is the false negatives observed for the *j*-th category. As shown, Pj, Rj, and Fj values increase with a rise in the algorithm prediction performance for a given category.

However, the accuracy of an algorithm depends on the different categories. Therefore, we adopted an *F* value defined as follows:

$$F=\frac{1}{n}\sum_{j=1}^{n}F_j, \tag{16}$$

where *n* is the number of category *j* in the training dataset.

Comparison based on pre-trained models

The accuracy of an algorithm varies and depends on the category. As a result, we established the following definition of an F value. By comparing their F scores, we evaluated **t**he benefits of the various pre-trained models considered for NER of Expressway Green Channel predictions. Suitability for using text data is a significant advantage of BiGRU-based hybrid deep learning networks. We obtained significant benefits from the pre-trained models by combining the BiGRU-CRF method with BERT, ALBERT, and RoBERTa.

The experimentation environment for BERT, ALBERT, and RoBERTa was provided by Google. QLoRA fine-tuning technique, a lightweight and efficient approach for training deep learning models, is employed to fine-tune specific parameters in the pre-trained model instead of retraining the entire model. Given the experiment's outcomes, our study comprises 12 Transformer layers, 768 hidden layers, and 12 heads of self-attention mechanism. With repeated experiments, we concluded that the recognition effect was the best when the super parameters were set, as shown in Table 3. These QLoRA fine-tuning parameters are based on general recommendations in named entity recognition tasks.

However, we identified the presence of some random selection biases in deep learning network training. The same samples were retrained three times independently using the same data handling techniques to avoid random biases in network training. The training dataset was used to train the independent deep learning network, expressed as Test 1, Test 2, and Test 3, displaying the performance of each pre-trained model in the test dataset in Table 4.

		Test 1			Test 2			Test 3		
Model	Entity	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
BERT	Cheating	93.42	93.42	93.42	98.63	94.74	96.64	98.63	94.74	96.64
	Mixed	100.00	99.85	99.92	100.00	100.00	100	100.00	100.00	100
	Out-listed	96.66	98.43	97.54	96.29	98.30	97.28	96.29	98.30	97.28
	Not fresh	92.06	92.06	92.06	89.23	92.06	90.62	89.23	92.06	90.62
	Frozen	82.11	76.52	79.22	83.61	77.27	80.31	83.61	77.27	80.31
	Spoilage	91.87	96.58	94.17	94.21	97.44	95.8	94.21	97.44	95.8
	Over-loading	99.32	98.65	98.99	99.55	99.21	99.38	99.55	99.21	99.38
	Deep-processing	83.19	85.34	84.26	84.42	85.63	85.02	84.42	85.63	85.02
ALBERT	Cheating	68.57	63.16	65.75	77.61	68.42	72.73	77.61	68.42	72.73
	Mixed	96.54	97.57	97.05	96.56	98.18	97.36	96.56	98.18	97.36
	Out-listed	86.15	91.90	88.93	87.69	92.16	89.87	87.69	92.16	89.87
	Not fresh	73.85	76.19	75	69.12	74.60	71.76	69.12	74.60	71.76
	Frozen	76.61	71.97	74.22	77.24	71.97	74.51	77.24	71.97	74.51
	Spoilage	75.71	90.60	82.49	72.03	88.03	79.23	72.03	88.03	79.23
	Over-loading	92.28	92.59	92.44	91.51	91.92	91.71	91.51	91.92	91.71
	Deep-processing	63.99	70.98	67.3	64.25	68.68	66.39	64.25	68.68	66.39
RoBERTa	Cheating	95.95	93.42	94.67	93.67	97.37	95.48	93.67	97.37	95.48
	Mixed	100.00	99.85	99.92	100.00	100.00	100	100.00	100.00	100
	Out-listed	97.30	99.08	98.19	97.16	98.56	97.86	97.16	98.56	97.86
	Not fresh	95.08	92.06	93.55	93.55	92.06	92.80	93.55	92.06	92.80
	Frozen	83.61	77.27	80.31	85.48	80.30	82.81	85.48	80.30	82.81
	Spoilage	92.62	96.58	94.56	91.06	95.73	93.33	91.06	95.73	93.33
	Over-loading	99.22	99.33	99.27	99.44	99.10	99.27	99.44	99.10	99.27
	Deep-processing	80.51	81.90	81.2	84.01	83.05	83.53	84.01	83.05	83.53

Table 4. Results based on BiGRU-CRF in three tests with different pretraining models (%).

Model	Test1 (%)	Test2 (%)	Test3 (%)	Mean value (%)	Standard deviation
BERT-BiGRU-CRF	92.45	93.13	93.13	92.90	0.3223
ALBERT-BiGRU-CRF	80.40	80.45	80.45	80.43	0.0224
RoBERTa-BiGRU-CRF	92.71	93.14	93.14	92.99	0.2009

Table 5. Comparison of F1 scores obtained by different models (%).

We calculated the mean value of F scores obtained by different models: BERT-BiGRU-CRF, ALBERT-BiGRU-CRF, and RoBERTa-BiGRU-CRF. Table 5 lists the F scores obtained by each pre-trained BERT, ALBERT, and RoBERTa model. Figure 6 displays the visual graphical presentation of the results in Table 5. Accordingly, the results provided a detailed evaluation of the advantages of the various pre-trained models.

It is apparent from Table 4 that the F1-score is associated with pre-trained models. This conclusion is consistent with the reported study by Sun et al.[19]. According to Table 4, the experimental results show that the precision of the RoBERTa-BiGRU-CRF model is 93.04%, the recall rate is 92.99%, and the F1-score is 92.99%, respectively. It is apparent from Table 5 that the RoBERTa-BiGRU-CRF model achieves better performance than baseline models. Specifically, the F1 value achieved the best of 93.14% by RoBERTa in three experiments. It is interesting that the mean F1 value of BERT-BiGRU-CRF and RoBERTa-BiGRU-CRF are fairly close as shown in Table 5. However, the lower Standard deviation was reported by RoBERTa model in experiments. We believe that RoBERTa-BiGRU-CRF is more stable over multiple training sessions. Overall, these results indicate that the RoBERTa pre-trained model outperforms other pre-trained models when BiGRU-CRF model are employed. Furthermore, due to the improvement of model parameters, bacth size and training corpus, these results suggest that RoBERTa model is a more effective solution than BERT and ALBERT, with the better effect on information extraction from Green Channel expressway textual records.

Comparison based on prediction models

According to previous studies, LSTM and GRU-based hybrid deep learning networks offer superior prediction accuracy than traditional machine learning algorithms. Then experimental analysis on LSTM and GRU-based models are presented for better performance in the deep learning module of Expressway Green Channel NER model. The F1 scores are evaluated for the benefits of the various prediction models.We compared our proposed model with BERT-GRU-CRF, BERT-LSTM-CRF, ALBERT-GRU-CRF, ALBERT-LSTM-CRF, RoBERTa-GRU-

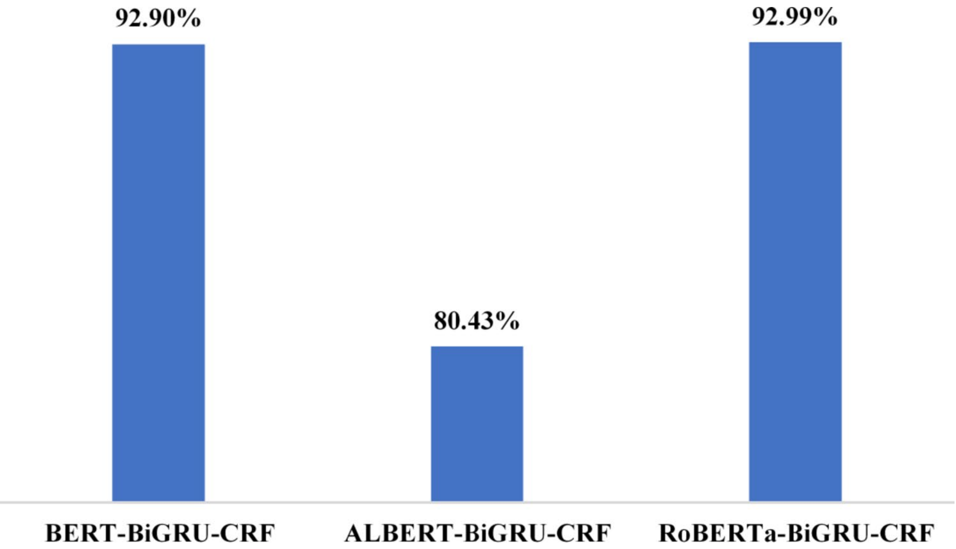

Fig. 6. Mean value of F with different pre-trained models.

Model	BERT-GRU-CRF (%)	BERT-LSTM-CRF (%)	BERT-LSTM-CRF (%)	ALBERT-LSTM-CRF (%)	RoBERTa-GRU-CRF (%)	RoBERTa-LSTM-CRF (%)
Cheating	0	0	35.00	3.60	93.24	89.33
Mixed	95.94	87.66	96.71	50.35	99.39	99.54
Out-listed	84.88	41.90	83.95	74.24	96.60	95.47
Not fresh	0	0	63.72	66.00	89.92	91.06
Frozen	65.10	60.16	69.02	53.11	81.10	81.42
Spoilage	60.44	18.12	54.62	40.72	94.07	87.18
Over-loading	51.03	24.79	86.29	86.30	99.27	98.87
Deep-processing	64.17	7.21	57.10	48.81	82.59	78.83
Mean value	52.70	29.98	68.30	52.89	92.02	90.21

Table 6. Comparison of different deep-learning-based models (%).

CRF, and RoBERTa-LSTM-CRF. As a result, Table 6 lists the F scores obtained by different deep-learning-based models.

In Table 6, from the comparison of the experimental results of LSTM and GRU, it can be seen that the GRU performs better than the LSTM base on the same structure. Base on GRU-CRF module, it can be seen that the RoBERTa-GRU-CRF model has highest F1 value of 92.02%. Base on LSTM-CRF module, it can be seen that the RoBERTa-LSTM-CRF model has highest F1 value of 90.21%. This verifies that RoBERTa pre-train module achieves the best performance, which was shown in our previous results.

In the comparison with RoBERTa-GRU-CRF and RoBERTa-BiGRU-CRF, the RoBERTa-BiGRU-CRF model achieves the higher overall score, and the F1 score is improved by 1%. In conclusion, the BiGRU network can better capture the context information of the serialized text, with stronger learning ability that is better than GRU or LSTM. The proposed method, namely RoBERTa-BiGRU-CRF model, introduces two features of pre-train and deep learning on the basis of the CRF module. From the overall point of view, the RoBERTa-BiGRU-CRF model achieves the best performance on the test dataset.

Ablation study

Further data analysis is required to determine how pre-trained and deep learning modules affect NER in Expressway Green Channel. Accordingly, the ablation study is evaluated according to the necessity and significance of each component for our proposed model. By eliminating a module, a control group is intended to demonstrate the necessity of that module. However, evidence has shown that the module has played a significant role when the performance is significantly reduced after deleting a module.

Text annotation, pre-trained language, BiGRU, and CRF modules comprise the NER of Expressway Green Channel based on text information. However, CRF and the text annotation modules are vital basic modules. Thus, we removed pre-trained and deep-learning modules to conduct ablation testing.

Table 7 lists the F scores of the different deep-learning-based prediction models obtained after removing the pretraining module, developed using BiLSTM-CRF and BiGRU-CRF. Table 8 lists the F scores obtained by the different models after removing the deep-learning-based prediction module, developed using BERT-CRF,

Model	BiLSTM-CRF			BiGRU-CRF		
	Precision	Recall	F1 (%)	Precision	Recall	F1 (%)
Cheating	0	0	0	0	0	0
Mixed	99.62%	92.26%	95.80	99.01%	91.58%	95.15
Out-listed	0	0	0	70.15%	1.08%	2.13
Not fresh	0	0	0	0	0	0
Frozen	0	0	0	0	0	0
Spoilage	0	0	0	0	0	0
Over-loading	99.49%	5.65%	10.69	96.39%	37.64%	54.14
Deep-processing	0	0	0	30.72%	0.01%	0.02
F	–	–	13.31	–	–	18.93

Table 7. The performance after removing the pretraining module.

Model	BERT-CRF			ALBERT-CRF			RoBERTa-CRF		
	Precision	Recall	F1 (%)	Precision	Recall	F1 (%)	Precision	Recall	F1 (%)
Cheating	82.56%	93.42%	87.65	37.62%	50.00%	42.94	96.10%	97.37%	96.73
Mixed	98.50%	99.85%	99.17	95.38%	97.26%	96.31	100.00%	100.00%	100
Out-listed	92.90%	97.52%	95.15	76.71%	86.54%	81.33	97.05%	98.95%	97.99
Not fresh	80.28%	90.48%	85.07	64.00%	76.19%	69.57	89.23%	92.06%	90.62
Frozen	84.80%	80.30%	82.49	64.75%	68.18%	66.42	82.11%	76.52%	79.22
Spoilage	88.19%	95.73%	91.80	66.90%	82.91%	74.05	91.13%	96.58%	93.78
Over-loading	97.67%	98.88%	98.27	87.03%	91.13%	89.04	99.89%	100.00%	99.94
Deep-processing	75.48%	79.60%	77.48	55.56%	66.09%	60.37	83.38%	85.06%	84.21
F	–	–	89.64	–	–	72.50	–	–	92.81

Table 8. The performance after removing the deep learning module.

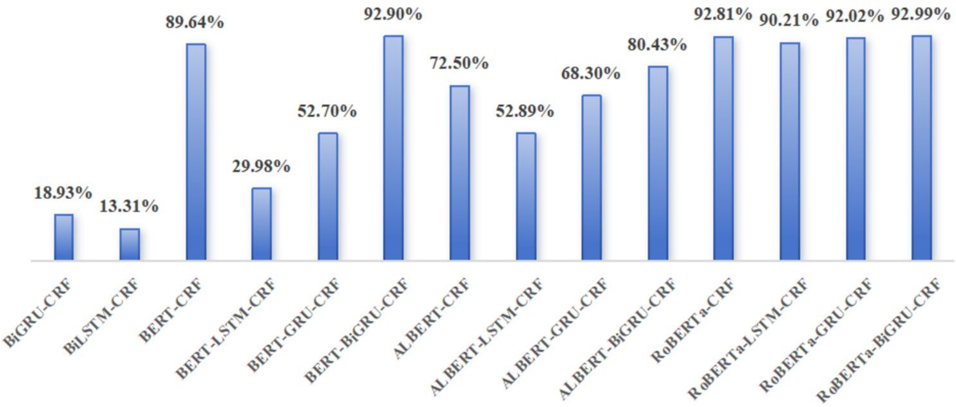

Fig. 7. Comparison of F scores obtained by models with the proposed method.

ALBERT-CRF, and RoBERTa-CRF. In addition, Fig. 8 presents a visual presentation of Tables 7 and 8 compares our proposed model with the results of different models.

Figure 7 shows test model performance using Expressway Green Channel dataset. It can be seen in Fig. 7 that the model performance was worse without pre-trained module, such as BiLSTM-CRF and BiGRU-CRF. Closer inspection from RoBERTa-CRF and RoBERTa-GRU-CRF shows that the introduction of a complex architecture was useless on improving performance. This result was shown when BERT-CRF, BERT-LSTM-CRF and BERT-GRU-CRF were compared. From the various experimental results, it can be found that the use of the pre-train task improves the final performance to a certain extent. In addition, the framework based on RoBERTa-BiGRU-CRF is effective with best performance, it can be used for Green Channel entity recognition task.

Discussion

The pre-trained models can capture meaningful features using the text data of the Expressway Green Channel. Assessing the significance of pre-trained language when working with text data is a crucial criterion for this study. Thus, our findings confirm that the textual features extracted by the pre-trained model substantially enhance the prediction capabilities of deep learning algorithms in the Expressway Green Channel NER tasks. The RoBERTa-BiGRU-CRF has the best effect in all aspects, which is the most remarkable finding from the data. The correlation between RoBERTa and BERT is worth mentioning because their results closely align with multiple tests. However, the RoBERTa offers a remarkable stability advantage with a lower standard deviation. This result is consistent with our earlier observations, showing that the RoBERTa improvement strategy is critical in our research. ALBERT had poor performance with Expressway Green Channel text data. As a result, the vectorization of text data successfully made use of the strengths of the more advanced RoBERTa and BERT pre-trained models. As a result, RoBERTa provides enough feature information for the pre-trained language model.

Importantly, GRUs are simplified based on LSTMs without sacrificing performance. According to the findings, the F score of GRU is slightly higher than LSTM's scores. The BERT model discovered the most obvious gap after pre-training. The BiGRU should theoretically outperform the GRU by collecting the contextual features in two directions. Initially, we thought that enhanced deep learning networks would help extract text features. In an unexpected finding, the RoBERTa-GRU-CRF model performed better than the ALBERT-BiGRU-CRF model in Green Channel expressway NER. Consequently, the pre-trained process is the critical constraint to the overall architecture after careful analysis. It is shown that the performance improvements is limited through advanced deep learning networks.

The ablation experiments are taken in our study as the same as other studies. The test was successful because removing each module revealed significant differences. Contrary to expectations, this study did not find a significant difference between RoBERTa-CRF and RoBERTa-BiGRU-CRF. However, the RoBERTa-BiGRU-CRF has a slightly higher F1 score than the RoBERTa-CRF. Nevertheless, we posit that the deep learning module can still be an essential component in the complex issue of Green Channel expressway NER.

In order to verify the effectiveness of this RoBERTa-GRU-CRF method, it is compared with other mainstream NER methods on Chinese in recent years. The specific results are shown in Table 9.

In Table 9, Qi et al. used Iterated Dilated Convolutional Neural Network (IDCNN) based on BiLSTM, and the F1 value was 86.10%. The experimental results from Chen et al. show that the F1-score of the FGNER (Fine-grained Geological Named Entity Recognition) model was 92.05%. Gao et al. propose a model using Multi-Task Learning and Biaffine Mechanism (MTL-BAM) base on BERT. The F1 value of XLNet-BilSTM-CRF model proposed by Sun et al. was close to 92.27%. Compared with the above model, our model has the best performance, the F1 value reaches 92.99%.

In our study, the precision of the RoBERTa-BiGRU-CRF model is 93.04%, the recall rate is 92.99%, and the F1-score is 92.99%, respectively, thus achieving better performance than baseline models. It can be seen from the table that the RoBERTa-BiGRU-CRF model is higher than the other entity recognition models. The results show that this method, which is superior to other methods, can effectively recognize named entities from green channel textual records on expressways.

The present results are meaningful in at least two major respects. The RoBERTa-BiGRU-CRF is the best-performing model in the Green Channel expressway NER. The pre-trained language model provides a significant contribution to performance improvement. These findings offer evidence in favor of processing and extracting textual information in Green Channel expressway. Our findings suggest that knowledge discovery of failure causes could be a big step forward for less cost in time and money.The proposed method can be effectively applied in practical scenarios to identify the perplexing agricultural products which not listed by the Green Channel, as well as address some frequently occurring misunderstandings for drivers. Using the proposed model, we can extract structured experiences for justifications of failure cases from unstructured electronic records of expressway toll stations. As a result, toll station inspectors can detect some high-frequency cheating behaviors as well as hidden fraudulent activities.

Although the proposed RoBERTa-BiGRU-CRF model achieves good results, the model is only trained on the specialized dataset in the Green Channel field, which limits the generalization ability of the model. Hence, further optimization is crucial regarding the representation and interpretation of knowledge.

Conclusions

Large heterogeneous electronic records of the Expressway Green Channel are created with information technology advancements. Information extraction of failure cases on these natural language e-records will enhance the policy implementation for both drivers and managers. This article presented an algorithmic model that utilizes the Expressway Green Channel dataset analytics and machine learning to recognize predefined entities. Moreover, our proposed solution is a heterogeneous deep learning architecture employing a pre-trained model to improve the accuracy of NER for the Expressway Green Channel. Our study is novel because few prior research is available on this subject. Our presented results can be directly utilized to advise drivers on

Model		Precision (%)	Recall (%)	F1 (%)
Chen et al.[8] (2021)	FGNER	95.73	89.26	92.05
Qi et al.[9] (2022)	IDCNN-BiLSTM	86.10	85.84	86.10
Gao et al.[13] (2022)	MTL-BAM	84.88	85.78	85.33
Sun et al.[19] (2022)	XLNet-BilSTM-CRF	92.80	91.74	92.27
Ours	RoBERTa-BiGRU-CRF	93.04	92.99	92.99

Table 9. Named entity recognition comparison experiment results (%).

making decisions. Overall, the proposed framework is an innovative and valuable work based on BIO labeling, pre-trained model, deep learning and CRF. As the basis for analyzing unqualified vehicles and the expressway toll-free vehicles of fresh agricultural products, these results add to the rapidly expanding the Expressway Green Channel text information field.

Our study established the Expressway Green Channel NER model to address the current shortage of unqualified vehicle cause analysis. The study's strengths included the in-depth analysis of the proposed framework's pre-trained and deep learning modules. We compared the BERT-BiGRU-CRF, ALBERT-BiGRU-CRF, and RoBERTa-BiGRU-CRF and verified that the RoBERTa-BiGRU-CRF model offered the most robust recognition effectiveness. In addition, we rigorously evaluated the influence of deep learning models by comparing the obtained F score. Our survey data from Shaanxi Province, China, showed that pretraining contributed significantly to the proposed framework. The results further demonstrated that the pre-trained model of RoBERTa best supported the performances of the deep-learning-based prediction algorithms.

This study is the first to investigate the Expressway Green Channel text information systematically. Our study provides a better understanding of the Expressway Green Channel data mining and sheds new light on unqualified vehicle cause analysis. However, the current research was limited to data from the Shaanxi Province of China and only considered the pre-trained models base on BERT. In future work, we plan to better improve the performance of the Green Channel NER model by combining external knowledge and enhance pre-trained model.

Data availability

The data used in this study are available from the corresponding author upon request.

Received: 12 January 2024; Accepted: 9 December 2024
Published online: 28 December 2024

References

1. Kong, L. S. et al. Spatial–temporal circulation pattern of fresh agricultural products based on green traffic data: a case study of Yunnan Province. *Transp. Res.* **8**(2), 87–95 (2022).
2. Liu, Y. et al. Prediction of fake toll-free vehicles based on historical traffic data. *J. Highway Transp. Res. Dev.* **15**(2), 92–102. https://doi.org/10.1061/JHTRCQ.0000775 (2021).
3. Lee, W. & Choi, J. Precursor-induced conditional random fields: connecting separate entities by induction for improved clinical named entity recognition. *BMC Med. Inf. Decis. Mak.* **19**(1), 132. https://doi.org/10.1186/s12911-019-0865-1 (2019).
4. Yi, F., Jiang, B., Wang, L. & Wu, J. J. Cybersecurity named entity recognition using multi-modal ensemble learning. *IEEE Access* **8**, 63214–63224. https://doi.org/10.1109/ACCESS.2020.2984582 (2020).
5. Furrer, L., Jancso, A., Colic, N. & Rinaldi, F. OGER plus plus: hybrid multi-type entity recognition. *J. Cheminform.* **11**(7), 3. https://doi.org/10.1186/s13321-018-0326-3 (2019).
6. Han, X. M. et al. MAF-CNER: A Chinese named entity recognition model based on multifeature adaptive fusion. *Complexity* **2021**, 6696064. https://doi.org/10.1155/2021/6696064 (2021).
7. Shi, X. et al. Extracting entities with attributes in clinical text via joint deep learning. *J. Am. Med. Inform. Assoc.* **26**(12), 1584–1591. https://doi.org/10.1093/jamia/ocz158 (2019).
8. Chen, T. Y. & Hu, Y. M. Entity relation extraction from electronic medical records based on improved annotation rules and BiLSTM-CRF. *Ann. Transl. Med.* **9**(18), 3828. https://doi.org/10.21037/atm-21-3828 (2021).
9. Qi, R. L., Lv, P. T., Zhang, Q. H. & Wu, M. Research on Chinese medical entity recognition based on multi-neural network fusion and improved tri-training algorithm. *Appl. Sci.-Basel* **12**(17), 8539. https://doi.org/10.3390/app12178539 (2022).
10. Kang, H. et al. A research toward Chinese named entity recognition based on transfer learning. *Int. J. Comput. Intell. Syst.* **16**(1), 56. https://doi.org/10.1007/s44196-023-00244-3 (2023).
11. Kim, D. et al. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* **7**, 73729–73740. https://doi.org/10.1109/ACCESS.2019.2920708 (2019).
12. Gong, L. J., Zhang, Z. F. & Chen, S. Q. Clinical named entity recognition from Chinese electronic medical records based on deep learning pretraining. *J. Healthcare Eng.* **2020**, 8829219. https://doi.org/10.1155/2020/8829219 (2020).
13. Gao, C. et al. A joint extraction model of entities and relations based on relation decomposition. *Int. J. Mach. Learn. Cybernet.* **13**(7), 1833–1845. https://doi.org/10.1007/s13042-021-01491-6 (2022).
14. Chen, S. Y. et al. Chinese fine-grained geological named entity recognition with rules and flat. *Earth Space Sci.* **9**(12), e2022002617. https://doi.org/10.1029/2022EA002617 (2022).
15. Fang, Q. & Li, Y. E. Chinese named entity recognition model based on multi-task learning. *Appl. Sci.-Basel* **13**(8), 4770. https://doi.org/10.3390/app13084770 (2023).
16. He, S. F., Sun, D. Q. & Wang, Z. Named entity recognition for Chinese marine text with knowledge-based self-attention. *Multimedia Tools Appl.* **81**(14), 19135–19149. https://doi.org/10.1007/s11042-020-10089-z (2021).
17. Wang, H., Zhou, L. K., Duan, J. Y. & He, L. Cross-lingual named entity recognition based on attention and adversarial training. *Appl. Sci.-Basel* **13**(4), 2548. https://doi.org/10.3390/app13042548 (2023).
18. He, B. & Zhang, J. R. An association rule mining method based on named entity recognition and text classification. *Arab. J. Sci. Eng.* **48**(2), 1503–1511. https://doi.org/10.1007/s13369-022-06870-x (2023).
19. Sun, J. L., Liu, Y. R., Cui, J. & He, H. D. Deep learning-based methods for natural hazard named entity recognition. *Sci. Rep.* **12**(1), 4598. https://doi.org/10.1038/s41598-022-08667-2 (2022).
20. Geng, R. S., Chen, Y. P., Huang, R. Z., Qin, Y. B. & Zheng, Q. H. Planarized sentence representation for nested named entity recognition. *Inf. Process. Manag.* **60**(4), 103352. https://doi.org/10.1016/j.ipm.2023.103352 (2023).
21. Li, H. J. et al. Named entity recognition for Chinese based on global pointer and adversarial training. *Sci. Rep.* **13**, 1. https://doi.org/10.1038/s41598-023-30355-y (2023).
22. Li, X., Yang, J. N., Liu, H. & Hu, P. J. HTLinker: a head-to-tail linker for nested named entity recognition. *Symmetry-Basel* **13**(9), 1596. https://doi.org/10.3390/sym13091596 (2021).
23. Gao, W. C. et al. Research on named entity recognition based on multi-task learning and biaffine mechanism. *Comput. Intell. Neurosci.* **2022**, 2687615. https://doi.org/10.1155/2022/2687615 (2022).
24. Zhang, Z. et al. Analyzing temporal complex events with large language models? A benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1588–1606 (Association for Computational Linguistics, 2024).

25. Chen, J., Tao, W. & Jin, Y. Modeling of the cause mechanism of unqualified toll-free vehicles of fresh agricultural products on expressway based on text mining. *Technol. Econ. Areas Commun.* **25**(6), 46–53 (2023).

Acknowledgements

We thank LetPub (https://www.letpub.com/) for linguistic assistance and pre-submission expert review.

Author contributions

Jiaona Chen: Conceptualization, Funding acquisition, Methodology, Project administration, Visualization, Writing—review & editing. Jing Zhang: Investigation, Data curation, Formal analysis, Software, Validation, Writing—review & editing. Weijun Tao: Investigation, Data curation, Formal analysis, Software, Validation, Writing—original draft. Yinli Jin: Project administration, Methodology, Supervision, Writing—review & editing. Heng Fan: Formal analysis, Writing—review & editing.

Funding

Jiaona Chen was supported by the National Natural Science Fund for Young Scholars [52002315]. Weijun Tao was supported by Postgraduate innovation and practice ability development fund of Xi'an Shiyou University [YCS22214248].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.C. or H.F.

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

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