IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2020, accepted May 22, 2020, date of publication May 29, 2020, date of current version June 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998486

Sequence Generation Network Based on Hierarchical Attention for Multi-Charge Prediction

KONGFAN ZHU*, BAOSEN MA*, TIANHUAN HUANG, ZEQIANG LI, HAOYANG MA, AND YUJUN LI, (Member, IEEE)

School of Information Science and Engineering, Shandong University, Qingdao 266200, China

Corresponding author: Yujun Li (liyujun@sdu.edu.cn)

*Kongfan Zhu and Baosen Ma contributed equally to this work.

This work was supported in part by the Key Research and Development Program of China under Grant 2018YFC0831000 and Grant 2017YFC0803400.

ABSTRACT The application of multi-label text classification in charge prediction aims at forecasting all kinds of charges related to the content of judgment documents according to the actual situation, which plays a vital role in the judgment of criminal cases. Existing classification algorithms have high accuracy for the single-charge prediction, but their accuracy for the multi-charge prediction is low. To solve this problem, in this paper we introduce a novel hierarchical nested attention structure model with relevant law article information to predict the multi-charge classification of legal judgment documents. By considering the correlation between different charges, the accuracy of multi-charge prediction is greatly improved. Experimental results on real-world datasets demonstrate that our proposed model achieves significant and consistent improvements over other state-of-the-art baselines.

INDEX TERMS Multi-charge prediction, hierarchical attention, sequence generation, logical correlation.

I. INTRODUCTION

In recent years, the task of charge prediction has attracted increasing attention. The purpose of this task is to predict the charges, law articles, terms of imprisonment, and other related information through given facts. Multi-charge prediction, as a representative sub-task of automatic charge prediction, plays an important role in the legal assistance system and can benefit many real-world applications. For example, it can provide legal experts with convenient reference information and thus improve their working efficiency. In addition, it can provide people who are unfamiliar with legal terminologies and complex procedures with legal consultation [1], [2].

Existing algorithms regard charge prediction as a single-label classification problem, by either adopting a K-nearest neighbor (KNN) [1], [3] as the classifier with shallow textual features or manually designing key factors for specific charges to help understand the text [4], which makes those works difficult to scale to multi-charge classification. In the single-charge prediction task, the single-charge model has a good prediction effect, but there are cases of ''one person with multiple charges'', resulting in the difficulty of extracting all the content features in the judgment documents. There are also works addressing a related task, finding the law articles that are involved in a given case. They often transform the multi-label classification problem into a multi-class classification task by only considering a fixed set of article combinations [5], which can only be applied to a small set of articles and does not fit real-world applications. Two improvements are proposed in the latest achievements: first a preliminary classification was performed and second a re-ranking method that deals with word-level and article-level features was used [5]. To some extent, these technologies have improved the experimental results, but they are heavily reliant on expert knowledge and extra textual analysis. Recent advances in neural networks have enabled us to jointly model charge prediction and relevant article extraction in a unified framework, where the latent correspondence from the fact description about a case to its related law articles and further to its charges can be explicitly addressed by a two-stack attention mechanism [6], [7]. However, these methods, which are used by setting a threshold, mostly ignore the logical correlation between different charges. Meanwhile, various parts of the text contribute differently to predicting different

The associate editor coordinating the review of this manuscript and approving it for publication was An-An Liu.

charges. Inspired by the tremendous success of the sequence-to-sequence model in machine translation, abstractive summarization, style transfer, and other domains, a sequence generation model consists of an encoder–decoder where the attention is proposed to generate labels sequentially, and thus predicts the next label based on its previously predicted labels [8]. In our proposed multi-charge prediction based on a sequence generation model, we employ the logical correlation between charges to capture the critical information relevant to some specific charges. By considering related factors, such as single-charge prediction, charge correlation, and relevant article extraction, multi-charge prediction could benefit from these related tasks on sequence generation models to achieve evident improvements.

The main problem of multi-charge classification is the explosive growth of output space [1]. Assuming that there are 20 tags, the output space has a power of 20. To deal with the label space with exponential complexity, it is necessary to mine the correlation between charges. For example, if a criminal commits the charge of ''smuggling'' and ''selling drugs'', the possibility of the offender committing the charge of ''detaining others to take drugs'' is also high, but the possibility of committing the charge of ''corruption'' or ''bribery'' is very low. Effective mining of the correlation between charges is the key to the success of multi-charge prediction.

In practice, there is a strong logical connection between the charges, such as ''theft'' and ''robbery'', or ''smuggling'', ''trafficking and transporting drugs'', and ''detaining other people to take drugs'', which have a high frequency of co-occurrence. In actual multi-charge prediction, the charge sequence is formed by sorting the charges, and the correlation information between the charges is integrated into the model to improve the prediction effect. In Table 1, we list the correlations between several charges [9].

TABLE 1. Correlations between several charges.

	Theft	Robbery	Murder	Rape
Theft	1	0.894	0.483	0.925
Robbery	0.894	1	0.581	0.937
Murder	0.483	0.581	1	0.667
Rape	0.925	0.937	0.667	1

In brief, our contributions are as follows. (1) We find that conventional multi-label classification algorithms are not suitable for the multi-charge classification, and we introduce a novel framework to consider the correlation between different charges to capture the critical information. (2) We propose a novel hierarchical nested attention structure model with relevant law article information to predict multi-charge classification of legal judgment documents. By considering the correlation between different charges, the accuracy of multi-charge prediction is greatly improved on the charge prediction datasets. (3) At the same time, our model can also improve the accuracy of the single-charge prediction.

The rest of this paper is organized as follows. In section II, we generalize the related works and further deduce the motivation of our model. Section III presents our structure based on this theory and method. In section IV, the experimental settings and model performance evaluation are presented. Finally, we present the conclusions and present future research directions.

II. RELATED WORK

For a long time, experts in the field of law have been studying how to achieve automatic charge prediction. Raghav and Krishna have applied quantitative methods to predict judgments by calculating numerical values for factual elements [10]. Katz attempted to extract efficient features from case annotations [11]. Liu *et al.* introduced mathematical models for charge prediction, such as linear models and the scheme of nearest neighbors [4]. These methods are usually mathematical or quantitative, and they only work with small datasets with few charges.

Past works have considered the multi-charge prediction as a special multi-class classification task that uses factual descriptions as inputs and outputs charge labels. Binary relevance (BR) transforms the multi-label classification task into multiple single-label classification problems by ignoring the correlations between labels [12]. Classifier chains (CC) transforms the multi-label classification task into a chain of binary classification problems and takes high-order label correlations into consideration [13]. Label powerset(LP) transforms the multi-label classification task into a multi-class problem with one multi-class classifier trained on all unique label combinations [14]. Based on the phrase classification method [10], KNN is used to classify criminal charges. However, the generalization ability of the KNN method is poor, and the word-level and phrase-level features extracted are too shallow to fully represent the text content of the charge fact description. It is impossible to obtain a sufficient basis to distinguish similar charges with nuances.

Lin *et al.* proposed a Chinese legal document labeling scheme by adding artificially labeled content as an aid to the machine learning model to improve the understanding of the case [5]. There are also scalability issues with this approach as the need to artificially design and annotate these determinants for each type of charge requires significant labor costs [6].

In the civil law system, some work has focused on determining the applicable legal provisions for a particular case. Phrase-based classification transforms this multi-charge problem into a multi-class classification problem by considering only a fixed set of articles [4], [15]. When considering a large set of legal articles, the number of possible combinations will increase exponentially, so this method cannot be extended to massive legal articles. The extremely multi-label text classification (XMTC) approach includes an extensible two-step classification method that first uses support vector machines (SVMs) to initially classify the article and then uses the word-level features and co-occurrence trends between the articles to sort the results [16]–[18].

Luo *et al.* proposed an SVM model to extract the top k candidate articles and article-side attention to better understand

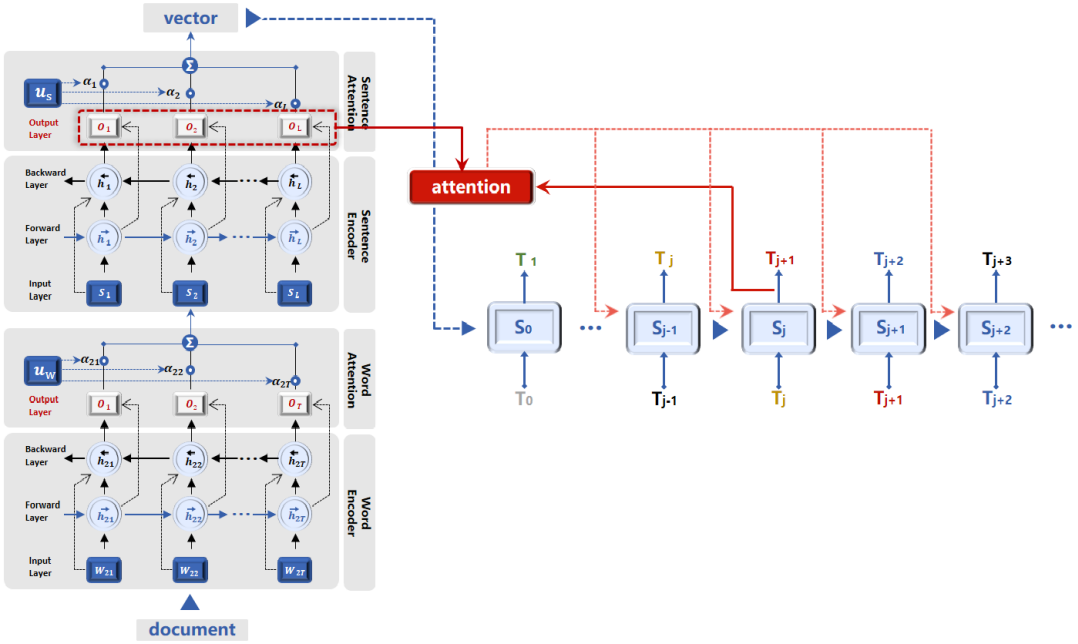


FIGURE 1. Overview of our proposed model. Our model consists of two parts: encoder transforms the criminal facts into critical information through word-level and sentence-level attention, and decoder transforms this information into charges sequentially with attention mechanism.

the texts, but the method relies on a threshold to predict relevant charges and does not consider the logical correlation between different charges [19], [20]. Compared with the generation model, this threshold model cannot reflect the characteristics of the training data itself and has limited capacity: this model has uncertainty reading the prior structure. Furthermore, this model can not catch the logical correlation between similar charges, and would not perform well when the number of classes increases and more similar charges appear [21].

By considering the correlations between labels, SGM [8] views the multi-label classification task as a sequence generation problem. It proposes an encoder–decoder with an attention mechanism structure to predict different labels. Experimental results show that this model outperforms other methods by a large margin. However, it is likely that this model would make a succession of wrong label predictions in the following time steps if the prediction is wrong at time step t. Meanwhile, it is not suitable for the representation of document-level input. SGM uses the generation model to generate labels sequentially, but it does not consider the relevant law article information to enhance the effect of multi-charge prediction. We adopt a relevant law article extractor as an auxiliary means to improve the prediction effect of our model. Concurrently, we adopt a completely different hierarchical nested attention mechanism, which can better capture relevant semantic information [22].

III. METHOD

A. OVERVIEW

An overall architecture of our proposed model is shown in Fig. 1. It consists of two parts: the encoder uses word-level attention to get the key information in the sentence, then uses sentence-level attention to get the key information about the facts of charges; and the decoder, with long short-term memory (LSTM) as the basic unit, which is used to decode the output vectors of the encoder and attention mechanism in charge prediction [23], where document refers to the document vector representation after processing, T_j denotes the jth charge and S_j denotes the hidden state of the decoder at time step j.

From the perspective of criminal facts, the charge prediction task can be modeled as finding an optimal charge sequence c^* that maximizes the conditional probability [24]:

$$P(c\,|f\,) = \prod_{i=1}^{n} P(c_i\,|c_1, c_2, \ldots, c_{i-1}, f\,), \tag{1}$$

where f is the fact of the judgment document, c denotes the charges contained in the judgment document, and c_i is a single charge.

Considering that a sentence is a combination of a series of words, and an article is a combination of a series of sentences, the text embedding problem can be transformed into a combination of words and sentence embedding problems by using hierarchical structure. In the following, we will present how to build the sentence-level and word-level vectors progressively from words by using word encoder and sentence encoder to solve the text embedding problem.

B. ENCODER

The structure of encoder is shown in Fig. 1. In the encoder, we adopt a hierarchical attention network. As the criminal facts text belongs to a long text level, we first perform the word-level attention operation on each sentence to achieve

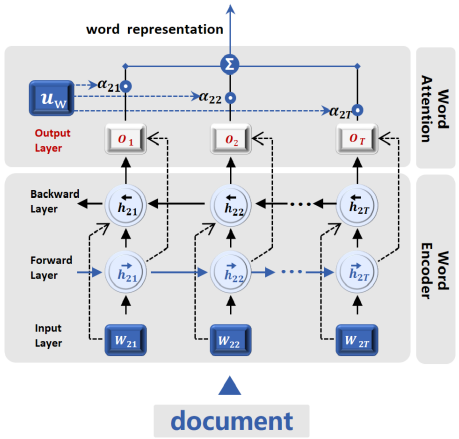


FIGURE 2. Word processing in the encoder.

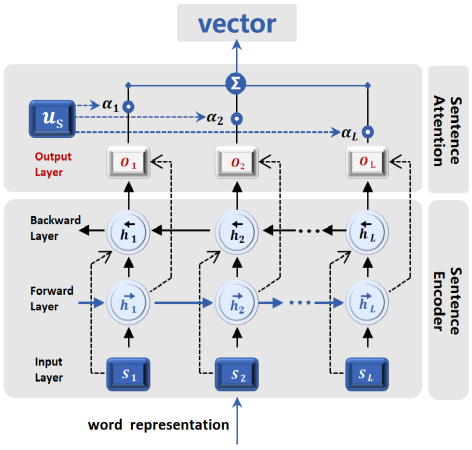


FIGURE 3. Sentence processing in the encoder.

feature extraction of each sentence. Then, the attention operation at the sentence-level is performed to obtain the feature representation of the entire text. On this basis, we perform the sentence-level attention operation. The key words and sentences in the criminal facts text can be obtained through the hierarchical attention operation [2], [7], [25].

1) WORD ATTENTION

Since different words in a sentence have different effects on the meaning of the entire sentence, it is helpful to introduce an attention mechanism to extract the words with important meanings and aggregate the representation of those informative words to form a sentence vector. We introduce the LSTM to represent the meaning of the sentence:

$$\begin{aligned} \overrightarrow{h}_t &= f(\overrightarrow{W} x_t + \overrightarrow{V}\,\overrightarrow{h}_{t-1} + \overrightarrow{b}), \\ \overleftarrow{h}_t &= f(\overleftarrow{W} x_t + \overleftarrow{V}\,\overleftarrow{h}_{t+1} + \overleftarrow{b}), \\ h_t &= [\overrightarrow{h}_t, \overleftarrow{h}_t], \end{aligned} \tag{2}$$

where W and V represent the weight matrices to be trained, b represents the bias vectors, x_t represents the input at time step t, h_t denotes the concatenated vector representation of forward hidden state $\overrightarrow{h}_t$ and backward hidden state; and $\overleftarrow{h}_t$ represents the hidden state at time step t.

The structure of word processing is shown in Fig. 2. First, we transform every word into a vector form by the embedding matrix. We introduce an attention mechanism to extract the words with important meanings and aggregate the representation of those informative words to form a sentence vector [26]:

$$\begin{aligned} u_{it} &= \tanh(W_w h_{it} + b_w), \\ \alpha_{it} &= \frac{\exp(u_{it}^{\mathrm{T}} U_w)}{\sum_t \exp(u_{it}^{\mathrm{T}} U_w)}, \\ s_i &= \textstyle\sum_t \alpha_{it} h_{it}, \end{aligned} \tag{3}$$

where W_w and U_w represent the word-level attention weight matrices, b_w represents the bias vector, and h_{it} represents the hidden state at time step t of the ith sentence. We first feed the word annotation h_{it} through a one-layer MLP to obtain u_{it} as a hidden representation of h_{it}, then we measure the importance of the word as the similarity of u_{it} with a word-level context vector u_w and obtain a normalized importance weight of it through a Softmax function [21]. Next, we compute the sentence vector s_i (we abuse the notation here) as a weighted sum of the word annotations based on the weights. The context vector u_w can be seen as a high-level representation of a fixed query "what is the informative word" over the words similarly to that used in memory networks. The word context vector u_w is randomly initialized and jointly learned during the training process [27]–[29].

2) SENTENCE ATTENTION

The structure of sentence processing is shown in Fig. 3. Given the sentence vectors s_i, we can obtain a document vector in a similar way. We use a bidirectional LSTM to encode the sentences:

$$\begin{aligned} \overrightarrow{h}_t &= f(\overrightarrow{W} s_i + \overrightarrow{V}\,\overrightarrow{h}_{t-1} + \overrightarrow{b}), \\ \overleftarrow{h}_t &= f(\overleftarrow{W} s_i + \overleftarrow{V}\,\overleftarrow{h}_{t+1} + \overleftarrow{b}), \\ h_t &= [\overrightarrow{h}_t, \overleftarrow{h}_t], \end{aligned} \tag{4}$$

where W and b represent the weight matrices and bias vectors, respectively; x_t represents the input at time step t; h_t denotes the concatenated vector representation of the forward hidden state $\overrightarrow{h}_t$ and backward hidden state; and $\overleftarrow{h}_t$ represents the hidden state at time step t. To reward sentences that are clues to correctly classifying a document, we again use an attention mechanism and introduce a sentence-level context vector u_s and use the vector to measure the importance of the sentences [30]. This yields

$$\begin{aligned} u_i &= \tanh(W_s h_i + b_s), \\ \alpha_i &= \frac{\exp(u_i^{\mathrm{T}} U_s)}{\sum_i \exp(u_i^{\mathrm{T}} U_s)}, \\ v &= \textstyle\sum_i \alpha_i h_i, \end{aligned} \tag{5}$$

where W_s and U_s represent the sentence-level attention weight matrices, b represents the bias vector, h_i represents the hidden state of the ith sentence, and v is the document vector that summarizes all the information of sentences in a document. Similarly, the sentence-level context vector can be randomly initialized and jointly learned during the training process [31].

C. DECODER

The decoder structure of our model is shown in Fig. 4, where T_j denotes the jth charge. In the decoder, we use the LSTM as the basic unit and draw on the attention mechanism in machine translation. On the one hand, it integrates the logical correlation between charges into the model; on the other hand, it strengthens the encoder–decoder information flow, thus completing the final multi-charge prediction [32]. The decoder reads each word vector sequentially and then copies the final hidden state to the decoder as the initial state. In the process of charge prediction, the decoder first reads the initial charge "sos" (start of the sentence), predicts the first charge described in the crime facts, then copies the charge to the second step as the input, and then predicts the second charge until the prediction result is the cut-off charge "eos" (end of the sentence) [33].

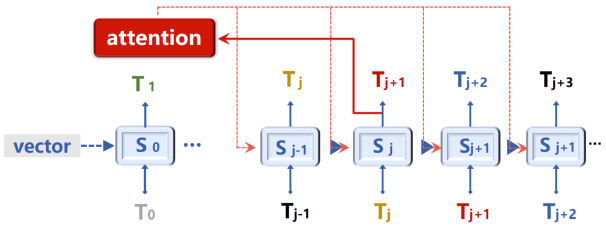


FIGURE 4. Decoder architecture.

The hidden state s_t of the decoder at time step t is computed as follows:

$$s_t = LSTM(s_{t-1}, [g(y_{t-1}); c_{t-1}]), \tag{6}$$

where $[g(y_{t-1}); c_{t-1}]$ represents the concatenation of the vectors $g(y_{t-1})$ and c_{t-1}, $g(y_{t-1})$ is the embedding of the label that has the highest probability under the distribution y_{t-1}. Here, c_{t-1} denotes the context vector at time step $t-1$ and y_{t-1} is the probability distribution over the label space L at time-step $t-1$ and is computed as follows:

$$\begin{aligned} O_t &= W_o f(W_d s_t + V_d c_t), \\ T_t &= Softmax(O_t), \end{aligned} \tag{7}$$

where O_t is the output of the LSTM cell and T_t is the probability of predicted charge at time step t. Here, W_o, W_d, and V_d are weight parameters, and f is a nonlinear activation function.

At the training stage, the loss function is the cross-entropy loss function. We employ the greedy search algorithm here. When the decoder predicts the end charge, the model stops predicting. The prediction paths ending with the "eos" are added to the candidate path set [34].

D. USING LAW ARTICLES

In the process of multi-charge prediction, we add an article extractor as an auxiliary means to improve the prediction effect of the model according to the content of the dataset. The first k law articles are selected by classifier, and then the feature vectors of these k articles are obtained by a neural network to represent semantic information, and the feature vectors are fed into the "attention" mechanism in Fig. 1. The extraction part of the law article is set according to the content of the dataset, which as mentioned in the experiment is legal information in the CJO dataset, and the extraction module of the law article can be added, but not in the CAIL dataset. We use the legal information in the data as the auxiliary means to predict the related charges, and then we combine the logical connection between the charges of criminal law to further improve the effect of charge prediction.

E. THE OUTPUT

To make the legal charge prediction, we first concatenate the document embedding and the aggregated article embedding, then use the full connection layer and Softmax layer to predict the classification charges. As the number of charges for each instance varies, we do not normalize the prediction probability [35], [36]. The loss function is given as follows:

$$J(\theta) = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} y_{ij}\log\hat{y_{ij}} + (1-y_{ij})\log(1-\hat{y_{ij}}), \tag{8}$$

where N is the number of samples, L is the number of labels, $\hat{y}_{ij} \in [0, 1]$ and $y_{ij} \in [0, 1]$ are the prediction probability and true values, respectively, for the ith sample and the jth label.

IV. EXPERIMENTS

Legal judgment documents are usually long texts with large amounts of words and data. First, there are 469 types of charges in criminal law. If the multi-charge prediction task is transformed into a two-class problem, 2^{469} new labels will be generated, which will cause huge manual processing costs, and significantly increase the complexity of the model. Using the sequence-based generation model, we add a startup label and end label, but the label space has not changed [37]. Considering the logical correlation between the charges, those charges that are less likely to occur at the same time are excluded.

In order to verify the effectiveness of our model on criminal prediction, we conducted experiments on datasets and compared our model with several state-of-the-art baselines [2].

For charge prediction, first we sort the charge sequence of each sample according to the frequency of the charges in the training dataset, where high-frequency charges are placed in the front. In addition, the "sos" and "eos" symbols are added to the head and tail of the charge sequence, respectively [5].

A. DATASETS CONSTRUCTION

We collect and construct two different legal judgment datasets: CJO and CAIL. CJO consists of criminal cases published by the Chinese government from China Judgement Online[1] and CAIL is another criminal case dataset of "Chinese AI and Law Challenge".[2] We selected 148,841 pieces of data from the CJO dataset, where each sample is divided into three parts: (1) criminal facts; (2) list of charges; (3) articles of law. There are 131 charges in the CJO dataset. We select 36,012 pieces of data from the CAIL dataset, each composed of the description of a case and the facts in the legal document. There are 106 charges in the CAIL dataset. It also includes the legal provisions involved in each case, the accused's conviction, and the length of the sentence.

B. BASELINES

We compare our proposed methods with the following baselines.

1) BINARY RELEVANCE

The basic idea of this algorithm is to decompose the multi-label classification tasks into Q independent binary classification problems, wherein each binary classification problem corresponds to a possible label in the label space. When the number of labels is large and the label density is low, class imbalance may occur in the binary classifier of each label [38].

2) CLASSIFIER CHAINS

The basic idea of this algorithm is to transform the multi-label learning problem into a chain of binary classification problems, where subsequent binary classifiers in the chain are built upon the predictions of preceding members. The disadvantage of this algorithm is that it loses the chance for parallel computing because it needs chain call to predict charges [13].

3) LABEL POWERSET

The basic idea of this algorithm is to transform the multi-label classification problem into the multi-class classification problem. Mapping 2^Q possible label sets to 2^Q natural numbers. The feature of this method is that the label set of label powerset (LP) prediction must already exist in the training set. It cannot generalize the label set that has never been seen before. As a result, the output space of this method is too large and the classification efficiency is low [14].

4) PREDICT CHARGES FOR CRIMINAL CASES WITH LEGAL BASIS

The predict charges for criminal cases with legal basis (fact_law) jointly models the charge prediction task and the relevant article extraction task [19]. We experiment with both the use of the relevant article information (fact_law) and the absence of the relevant article information (fact_wo).

[1]http://wenshu.court.gov.cn.
[2]http://cail.cipsc.org.cn/index.html.

5) HIERARCHICAL ATTENTION NETWORKS

Hierarchical attention networks (HAN) uses two levels of attention mechanism applied at the word-level and sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation [22].

C. EXPERIMENTAL SETTINGS

For the two datasets described previously, as the documents are well-structured and human-annotated, we can easily extract fact descriptions, applicable law articles, charges, and terms of penalty from each document using regular expressions.

For all models and baselines, we use Adam [39] as the optimizer, and set the learning rate 0.001, the dropout rate [40] 0.5 and the batch size 32. Since the case documents are written in Chinese with no spacings between words, we employ word segmentation. Afterward, we adopt the Skip–Gram model to pre-train the word embeddings on these case documents, with embedding size set 100 and frequency threshold set 25.

According to the statistics of the charges in the CAIL dataset, we decide that the frequency of single charges should be at least 50. In the CJO dataset, we set the minimum number of charges 80, and we sift the charges whose frequency below 80 out. Then we set the charge name dictionary according to the frequency.

The law needs to be textualized as input: we set the text length of the CJO 300 and the law of CJO 500, the text length of CAIL is set 400 and law remains unchanged. The number of certain charges related to the data is too small, we set the CJO's text length 300, the CJO's law 500, and the CAIL's text length 400.

D. EVALUATION METRICS

Following the previous work, we adopt hamming loss and macro-F1 score as our main evaluation metrics for the performance comparison, because both are widely used evaluation methods for multi-label classification problems [19]. Hamming loss is used to calculate the accuracy of the multi-label classification model:

$$Hammingloss = \frac{1}{N}\sum_{i=1}^{N}\frac{XOR(y_{ij}, \hat{y}_{ij})}{L}, \tag{9}$$

where N is the number of samples, L is the number of charges, $\hat{y}_{ij}$ is the true value of the jth component in the ith prediction result, $\hat{y}_{ij}$ is the predicted value of the jth component in the ith prediction result, and XOR is the "exclusive OR" operation. Here, we employ accuracy, macro-precision, macro-recall, and macro-F1 as our evaluation metrics [2], [8], all the formulas are defined as follows:

$$Accuracy_j = \frac{TP_j + TN_j}{TP_j + TN_j + FP_j + FN_j},$$

$$Precision_j = \frac{TP_j}{TP_j + FP_j},$$
$$Recall_j = \frac{TP_j}{TP_j + FN_j},$$
$$Precision_{macro} = \frac{1}{L}\sum_{j=1}^{L} Precision_j,$$
$$Recall_{macro} = \frac{1}{L}\sum_{j=1}^{L} Recall_j,$$
$$F1_{macro} = \frac{2 * Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}}, \quad (10)$$

where TP_j, TN_j, FP_j, and FN_j represent the number of true positive, true negative, false positive, and false negative test samples with respect to the jth charge, respectively.

E. RESULTS AND ANALYSIS ON MULTI-CHARGE PREDICTION

The performance comparison with the previous research work is demonstrated in Tables 2, 3, and 4. As shown in Fig. 5 and Fig. 6, the precision and loss of our model in the validation set in the training process is about 90%. Meanwhile, Fig. 7 shows that the Hamming loss of the CAIL dataset remains very low. Almost all existing methods perform poorly under the Macro-F1 metric, which shows that they do not effectively combine the law article information with the nesting of the attention structure. Conversely, our model achieves promising improvements, demonstrating the robustness and effectiveness of our method. By adding a law extractor, the law text encoder, introducing criminal law information improves the prediction effect of the model. However, without introducing the correlation information between labels, the multi-charge prediction effect is not as good as that of single-charge prediction.

TABLE 2. Comparison between our model and all the baseline models on the CJO test set (without law article information).

Model	Hamming Loss	Precision	Recall	F1
BR	0.012	0.823	0.312	0.453
CC	0.011	0.802	0.351	0.488
LP	0.014	0.556	0.531	0.543
Fact_law_wo	**0.008**	**0.880**	0.536	0.667
Our model_wo	0.009	0.853	**0.839**	**0.846**

TABLE 3. Comparison between our model (with law article information) and all the baseline models on the CJO test set.

Model	Hamming Loss	Precision	Recall	F1
BR	0.012	0.644	0.648	0.646
CC	0.011	0.657	0.651	0.654
LP	0.014	0.662	0.608	0.634
Fact_law	0.004	**0.955**	0.753	0.831
Our Model	**0.003**	0.939	**0.934**	**0.937**

TABLE 4. Comparison between our model (with law article information) and all the baseline models on the CAIL test set.

Model	Hamming Loss	Precision	Recall	F1
BR	0.011	0.879	0.501	0.639
CC	0.010	0.853	0.558	0.675
LP	0.012	0.690	0.669	0.679
HAN	0.006	0.927	0.747	0.827
Our Model	**0.005**	**0.931**	**0.929**	**0.930**

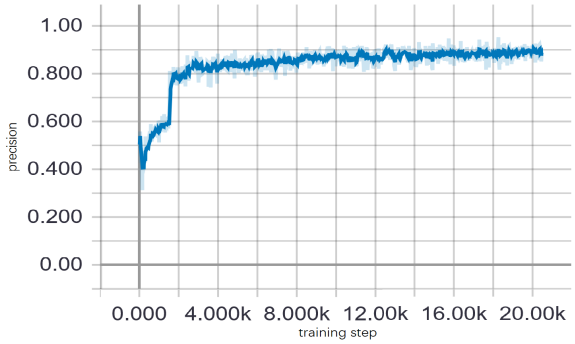


FIGURE 5. Precision in the training process.

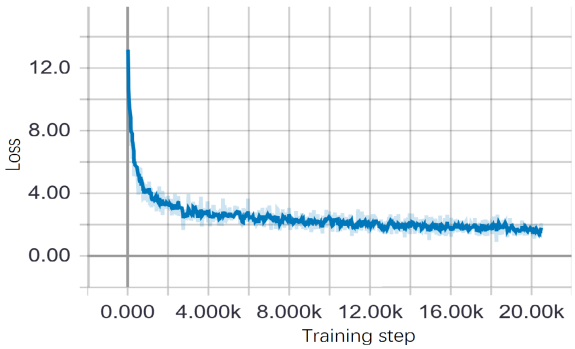


FIGURE 6. Loss in the training process.

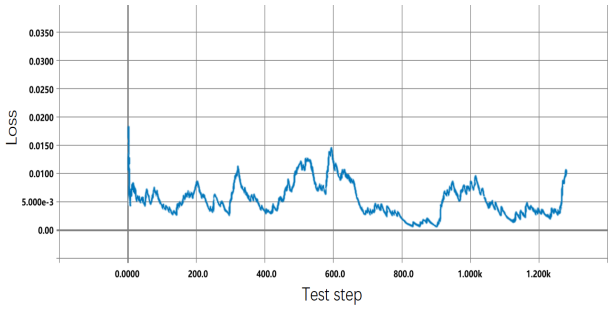


FIGURE 7. Hamming loss in the test process on the CAIL dataset.

F. RESULTS AND ANALYSIS ON SINGLE-CHARGE PREDICTION

Taking the single-charge prediction as a special case of the multi-charge prediction, our model can also be used to predict single charges. In the process of single-charge prediction, the decoder first reads the initial charge "sos," predicts the related charge described in the crime facts, then takes this charge as the input, and the prediction result is the cut-off charge "eos."

We can observe in Table 5 that our model also outperforms all the baselines. As shown in Fig. 8, the precision of our model in the training process is approximately 100%. Experimental results show that our model also has good robustness on single-charge prediction.

TABLE 5. Single-charge comparison between our model and baseline models on the CAIL test set.

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
LSTM	0.824	0.776	0.756	0.757
HAN	0.853	0.821	0.800	0.801
Our model	**0.880**	**0.832**	**0.819**	**0.820**

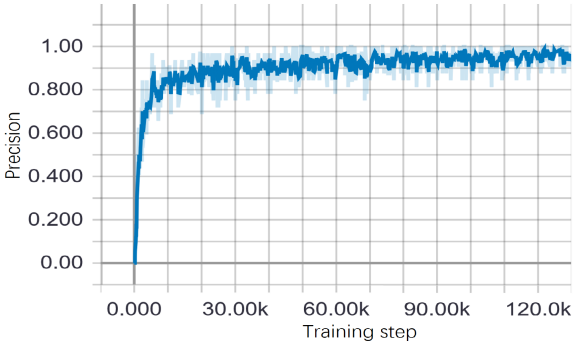


FIGURE 8. Precision in the training process.

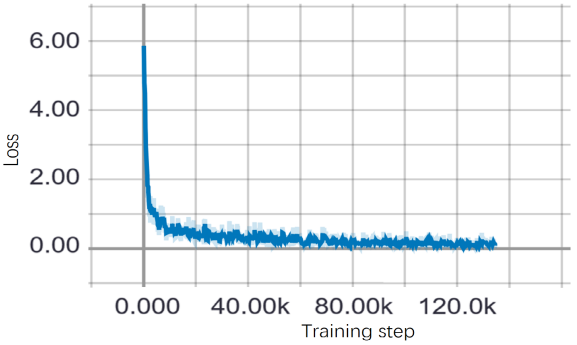


FIGURE 9. Loss in the training process.

G. CASE STUDY

In this section, we choose a representative case to show that the attention module helps improve the accuracy of multi-charge prediction in Fig. 10. In this case, the defendant was convicted of "obstruction of official duties" and "provocation." Because cases involve specific amounts of money and damaged goods, it is often difficult to decide whether a case should be judged as an "obstruction of public affairs" or a "provocation" because both are related to violence. An important feature of both is that the scenario in the case is the public goods of the administrative department, which will be given a higher weight in attention.

经 审理 查明，2016 年 5 月 31 日 14 时许 ，被告人 孙某某 酒后 来到 靖宇县 赤松乡 赤松村 村委会 村部，无故 将 村部 的 窗户 玻璃 以及 村干部 尹某某 停放 在 村部 院内 本田 思迪 轿车 砸坏。经 靖宇县 价格 认证 中心 鉴定：赤松村 村委会 村部 被 砸碎 的 16 块 玻璃 价值 人民币 2668 元，被 砸 的 本田 思迪 轿车 风挡 玻璃 及 车损 价值 人民币 4470 元，总 损失 价值 人民币 7138 元。
妨害公务罪#寻衅滋事罪

After trial, it was found that at about 14:00 on May 31, 2016, the defendant Sun Moumou came to the village committee of Chisong Village, Chisong Township, Jingyu County, and broke the window glass of the village and the village cadre Yin Mou-mou's Honda city car parked in the village yard without reason. According to the appraisal of Jingyu Price Certification Center, the 16 pieces of glass smashed in the village of Chisong Village Committee are worth RMB 2668 yuan, the windshield and car damage value of the smashed Honda city car is RMB 4470 yuan, and the total loss value is RMB 7138 yuan. The above facts are supported by the following evidences:
crime of disrupting public service # Crime of provoking trouble

FIGURE 10. Visualization of the attention mechanism.

Thus, we believe that the attribute is essential in the charge prediction of this case. In addition, we visualize the heat map of this case when predicting the attribute intentional injury. Words with deeper background color have higher attention weights. In Fig. 10, we observe that the attention mechanism can capture key patterns and semantics relevant to the current attribute.

V. CONCLUSION

In this work, we have focused on the task of multi-charge prediction according to the fact descriptions of criminal cases. To address the problem of predicting countless and confusing charges, we have introduced a novel hierarchical nested attention structure to predict multiple charges of legal judgment documents. Specifically, our model learns the hierarchical nested attention structure and legal judgment fact representation jointly by utilizing an attribute-based attention mechanism and logical correlation. Experimental results on CAIL and CJO datasets have shown that our model outperforms significantly all baselines and conventional multi-label classification models.

In future, we will explore the following directions. (1) To verify the generalization ability of the model, we will use our model in similar task in other languages. (2) We will explore more complicated legal judgment cases, such as multiple defendants and charges. Thus, it is challenging to handle this general form of charge prediction. (3) We will explore graph embedding with adversarial training methods to investigate the effectiveness of multi-charge prediction [41], [42]. (4) We will explore how to incorporate task-sensitive features to improve the performance of multi-charge prediction [43], [44].

In particular, if designed properly, transfer learning can provide encouraging results [45]–[48]. Meanwhile, we expect that all kinds of high-performance language models can be applied to the multi-charge prediction in the future.

REFERENCES

[1] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[2] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 487–498.

[3] C.-L. Liu, C.-T. Chang, and J.-H. Ho, "Case instance generation and refinement for case-based criminal summary judgments in Chinese," *J. Inf. Sci. Eng.*, vol. 20, no. 4, pp. 783–800, Jul. 2004.

[4] C.-L. Liu and C.-D. Hsieh, "Exploring phrase-based classification of judicial documents for criminal charges in Chinese," in *Proc. 16th Int. Symp. Found. Intell. Syst. (ISMIS)*, Bari, Italy, Sep. 2006, pp. 681–690.

[5] W.-C. Lin, T.-T. Kuo, T.-J. Chang, C.-A. Yen, C.-J. Chen, and S.-D. Lin, "Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction," in *Proc. ROCLING*, 2012, p. 140.

[6] C.-L. Liu and T.-M. Liao, "Classifying criminal charges in Chinese for Web-based legal services," in *Proc. Asia–Pacific Web Conf. Web Technol. Res. Develop.* Springer, 2005, pp. 64–75.

[7] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2377–2383.

[8] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: Sequence generation model for multi-label classification," 2018, *arXiv:1806.04822*. [Online]. Available: http://arxiv.org/abs/1806.04822

[9] C. Kim, K. Kyu, J. In, D. Park, and C. Hyun, "Analysis of the spatial characteristics of the five major crime using GIS," in *Proc. Collection Academic Conf. Korean Soc. Surveying*, 2014, pp. 273–275.

[10] K. Raghav, P. K. Reddy, and V. B. Reddy, "Analyzing the extraction of relevant legal judgments using paragraph-level and citation information," in *Proc. Artif. Intell. Justice (AI4JC)*, 2016, p. 30.

[11] D. M. Katz, M. J. Bommarito, and J. Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*. Rochester, NY, USA: Social Science Electronic Publishing, 2017.

[12] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[13] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, p. 333, Dec. 2011.

[14] Z. F. Salih and S. Tiun, "Term expansion and powerlabel set for multi-label hierarchical on short document classification," *Int. J. Appl. Eng. Res.*, vol. 13, no. 1, pp. 539–544, 2018.

[15] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1372–1383, May 2020.

[16] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," 2018, *arXiv:1811.01727*. [Online]. Available: http://arxiv.org/abs/1811.01727

[17] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[18] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, and K.-K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9280–9293, Dec. 2019.

[19] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," 2017, *arXiv:1707.09168*. [Online]. Available: http://arxiv.org/abs/1707.09168

[20] Z. Gao, H. Xue, and S. Wan, "Multiple discrimination and pairwise CNN for view-based 3D object retrieval," *Neural Netw.*, vol. 125, pp. 290–302, May 2020.

[21] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, Feb. 2010.

[22] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] R. Keown, "Mathematical models for legal prediction," *Computer/LJ*, vol. 2, no. 1, p. 829, 1980.

[25] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi, "Significantly fast and robust fuzzy C-Means clustering algorithm based on morphological reconstruction and membership filtering," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3027–3041, Oct. 2018.

[26] S. Baker and A. Korhonen, "Initializing neural networks for hierarchical multi-label text classification," in *Proc. BioNLP*, 2017, pp. 307–315.

[27] H. Ma, Y. Li, X. Ji, J. Han, and Z. Li, "MsCoa: Multi-step co-attention model for multi-label classification," *IEEE Access*, vol. 7, pp. 109635–109645, 2019.

[28] G. Kurata, B. Xiang, and B. Zhou, "Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2016, pp. 521–526.

[29] Y. Tagami, "AnnexML: Approximate nearest neighbor search for extreme multi-label classification," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 455–464.

[30] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari, "LSHTC: A benchmark for large-scale text classification," 2015, *arXiv:1503.08581*. [Online]. Available: http://arxiv.org/abs/1503.08581

[31] S. Gopal and Y. Yang, "Multilabel classification with meta-level features," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2010, pp. 315–322.

[32] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[34] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 681–687.

[35] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[36] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[37] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. NIPS*, 2014, *arXiv:1409.3215*. [Online]. Available: https://arxiv.org/abs/1409.3215

[38] Z. Liu, J. Wang, G. Liu, and L. Zhang, "Discriminative low-rank preserving projection for dimensionality reduction," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105768.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] S. Pan, R. Hu, S.-F. Fung, G. Long, J. Jiang, and C. Zhang, "Learning graph embedding with adversarial training methods," 2019, *arXiv:1901.01250*. [Online]. Available: http://arxiv.org/abs/1901.01250

[42] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019, *arXiv:1901.00596*. [Online]. Available: http://arxiv.org/abs/1901.00596

[43] S. Pan, J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Joint structure feature exploration and regularization for multi-task graph classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 715–728, Mar. 2016.

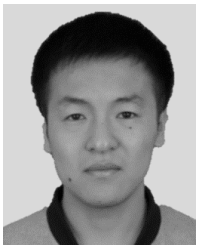
[44] S. Pan, J. Wu, X. Zhu, G. Long, and C. Zhang, "Task sensitive feature exploration and learning for multitask graph classification," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 744–758, Mar. 2017.

[45] X. Ben, P. Zhang, R. Yan, M. Yang, and G. Ge, "Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation," *Neural Comput. Appl.*, vol. 27, no. 8, pp. 2629–2646, 2016.

[46] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognit.*, vol. 90, pp. 87–98, Jun. 2019.

[47] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 734–747, Mar. 2020, doi: 10.1109/TCSVT.2019.2893736.
[48] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, Jun. 2019.

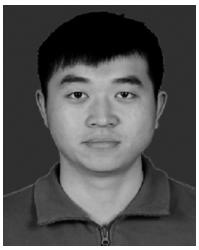
KONGFAN ZHU received the B.S. degree in computer science from Heilongjiang University, Harbin, China. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. He has been engaged in the informatization research work in the judicial field for a long time. He has a strong background in judicial informatization and big data research. His current research interests include machine learning and natural language processing.



BAOSEN MA received the B.S. degree in electronic information science and technology from the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China, in 2017. He is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Shandong University, Qingdao, China. His current research interests include machine learning, data mining, and natural language processing.

TIANHUAN HUANG received the B.E. degree in electronic information engineering from the School of Physical Science and Technology, Nanjing Normal University, Nanjing, China, in 2018. She is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. Her current research interests include machine learning, image processing, and deep learning.

ZEQIANG LI received the B.S. degree in electronic and information engineering from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2017. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. His current research interests include machine learning, data mining, natural language processing, and sentiment analysis.



HAOYANG MA received the B.S. degree in applied physics from the School of Physical Science, Qingdao University, Qingdao, China, in 2017. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao. His current research interests include machine learning, data mining, multi-label learning, and few-shot learning.

YUJUN LI (Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2001. He is currently a Full Professor with the Department of Information Science and Engineering, Shandong University, Qingdao, China. His current research interests include deep learning, natural language processing, multi-label learning, and sentiment analysis.