Computer Science and Information Systems 17(3):795–817 https://doi.org/10.2298/CSIS190903021Y

VDRF: Sensing the Defect Information to Risk Level of Vehicle Recall based on Bert Communication Model

Xindong You[1], Jiangwei Ma[1], Yuwen Zhang[1], Xueqiang Lv[1,*], and Junmei Han[2]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University, Beijing, 100101, China
[2] Laboratory of Complex systems, Institute of Systems Engineering. AMS. PLA, Beijing, 100029 China
*Corresponding Author lxq@bistu.edu.cn

Abstract. The recall of defective automobile products is one of the important measures to promote the quality of product quality and protect consumers' pyhsical safety and property security. In order to assess the risk level of defect cases, automobile recall management experts need to analyze and discuss the defect information by personal. A risk level prediction method based on language pre-training Bert model is proposed in this paper, which can transform the defect information into rick level of the vehicle and then predict vehicle recall automatically, in which a seq2seq model is proposed to multi-label the vehicle complaint data. The outputs of the seq2seq model combined with other static and dynamic information are used as the input of the Bert communication model. Substantial comparative experiments of different feature combinations on different methods show that the proposed VDRF method achieves F1 value with 79% in vehicle recall risk prediction, which outperforms the traditional method.

Keywords: Bert communication model, defect information transforming, multi-label classification, risk level prediction.

1. Introduction

With the continuous development of the vehicle industry, vehicles have become a necessity in people's lives. A large number of consumer complaints are collected in the vehicle quality defect complaint system, which named the defect information collection system of Defective Product Administrative Center [1]. A variety of problems or failures often occur during using the automobiles. Some of these problems are caused by improper operation or other external reasons in the process of using, and have nothing to do with the automotive products themselves. Another part of the problem is usually due to negligence in the production and design process of automobiles. These products have their own design defects, because of the particularity of the automobiles, these defects will threaten the safety of consumers' lives and property to a certain extent. Defects in vehicles can cause bodily harm and sometimes fatal consequences. Moreover, defects in automotive products can have a devastating impact on the sales and reputation of automakers, especially in the social media era. In order to avoid this risk, Europe and the United States established their own defective vehicle recall system. The status quo has formed a complete recall system of automotive products. With the continuous development of China's automobile industry, the corresponding defective automobile product recall system is gradually improving. Automobile consumers are more and more inclined to use the Internet platform to release vehicle defect information. Early detection of defects not only protects consumers from economic loss, but also mitigates the financial loss of manufacturers. In the process of defective vehicle recall, defect information is an important basis to judge the risk of vehicle recall. In order to assess the risk level of defect cases, automobile recall management experts need to analyze and discuss the defect information submitted by consumers to determine whether to carry out relevant recall work, which takes a lot of time and energy. Employing natural language processing technology and in-depth learning technology to process and analyze the defect information can help defect recall managers better analyze and assess the severity of automobile defects.

In order to evaluate the severity of automobile defects, the risk level prediction of automobile defective product recall is investigated fully in this paper, and a risk prediction model based on language pre-training Bert communication model (VDRF) is proposed. The proposed VDRF communication model can sensing the defect related information into the risk level of vehicle recall automatically. Firstly, the original data is preprocessed and a data set of automobile defect cases with a certain scale is constructed. Thereafter, multi-dimensional features are extracted, such as static features, dynamic features and fault semantic features. Finally, the extracted different combinations of features are used to predict the recall risk level of the vehicle.

As a whole, the main contributions of this paper are listed as follows:

(1) Two vehicle complaint datasets are constructed through utilizing web crawler technology, in which all kinds of complaints in the process of vehicle recall are contained.

(2) A Seq2seq neural network model is firstly employed to solve the multi-label classification on vehicle complaint data, in which the defect label features and defect label distribution are added to the basic seq2seq model, which makes the model more suitable for multi-label classification of vehicle complaint data.

(3) The pre-training language model Bert model is used to predict the risk lev-el of vehicle recall. Static feature, dynamic feature and fault semantic feature are extracted to classify the risk level, so that the semantic information in fault description can be better captured.

(4) Substantial comparative experiments of different feature combinations and different methods are conducted, which show that the proposed method achieves F1 value with 79% in vehicle recall risk prediction, which outperforms the traditional method.

2. Relate Work

Multi-label classification of the defect information is the preorder of the risk level prediction, and is the important part of this paper. Therefore, related work of multi-label classification and risk level prediction are investigated in this section.

Multi-label classification mainly includes three types of solutions. They are problem transformation methods, algorithm adaptation methods and neural network-based methods. The idea of problem transformation is to transform multi-label problem into single-label classification problem in some way, a mature single label classification method is used to solve the problem. Binary Reliance (BR) algorithm proposed by Boutell [2] transforms each label into a single label classification problem, which is independent of each other. The disadvantage of this method is that the relationship between labels is ignored. Similar algorithms include LIFT algorithm [3], Label Powerset (LP) algorithm [4], and Classifier Chain (CC) algorithm [5].

The algorithm adapts to multi-label data after modifying and extending the traditional single-label classification algorithm. Clare [6] extends the definition of information entropy to multi-label problem, and then uses improved decision tree algorithm to classify multi-label [7]. Elisseeff [8] proposes Rank-SVM algorithm by introducing loss function to support vector machine (SVM). Zhang and Zhou [9] proposed an improved ML-KNN algorithm based on k-nearest neighbor algorithm to solve the multi-label classification problem. Li [10] proposed a new joint learning algorithm, which propagates the feedback of the current label to the classifier of the subsequent label, and achieves good results in text multi-label classification.

Neural network models are applied to multi-label learning tasks recently. Zhang and Zhou [11] proposed BP-MLL model, which uses a new loss function in the fully connected neural network. Experiments show that the neural network model can capture the characteristics of multi-label tasks. Chen [12] uses a com-bination of CNN and RNN to represent the semantic information of the text and the higher-order features between the labels. Baker [13] maps the rows of co-occurrence labels to initialize the final hidden layer of the CNN, which can improve the performance of the model. Yang [14] claimed that multi-label classification task should be regarded as sequence generation problem. They use a new sequence generation model with a new decoder structure to solve the multi-label classification problem, and achieved good results.

In the research of automobile defect recall prediction, Zhang [16] proposed a new method to predict automobile recall risk based on the content published by users in the forum. For defective vehicles, before manufacturers and government agencies take investigative action, vehicle forums on the Internet typically dis-play user-posted content containing features of a defective vehicle. Through statistical analysis, it is found that there are overlaps between these contents and the official recall notices. It is of great significance for vehicle recall work to study the use of various machine learning algorithms to predict the risk of vehicle recall using defect features. Yang Shuanglong [17] collects the complaint data of various automobile platforms on the Internet through a large number of automated ways, and uses data mining methods to carry out risk recall early warning research on automobile products. It mainly includes automatic collection and pretreatment of automobile complaint data, text classification based on automobile complaint data, and early warning of automobile recall risk based on complaint data. Jiang Cuiqing [18] and others need a lot of manual labeling for the classification process in the research of automobile defect discovery, and that the classification of defective contents according to product components is not completely applicable. Based on Chinese social media, a framework of automobile defect recognition and automobile defect feature set in constructed in this paper, studies the method of automobile product defect classification using semi-supervised learning algorithm and the subject modeling of automobile product defect using LDA, and achieves good results. There are also some researches in this field abroad. Abrabhams [19] proposed a framework for automobile product defect detection based on the relevant information published by users on auto-mobile forums. In the framework, firstly, the relevant feature information is mined by text mining technology, and then a regression model for automobile product defect detection is constructed. Abrabhams [20] constructs a text mining model that can identify the auto parts involved in the user's post content. A binary classifier according to the name of the forum sub-module published by the post as the tag of the post is constructed in this paper, which classifies the content of the forum post according to the auto parts involved [21].

3. VDRF: Prediction of Risk Level of Vehicle Recall based on Defect Information

3.1. Model Architecture of VDRF

An overview of our proposed model is Figure 1. Firstly, we construct the automobile domain dictionary and the automobile defect label library according to the data on the Internet. Then we expand the automobile defect label library by using the automobile domain dictionary, and get the synonymous description of the automobile defect label library. According to the automobile defect label library, we classify the data of automobile complaints and get the defect label. Finally, we use Bert to predict the risk level of automobile recall based on the static and dynamic features, defect labels and defect severity levels extracted from automobile defect data.

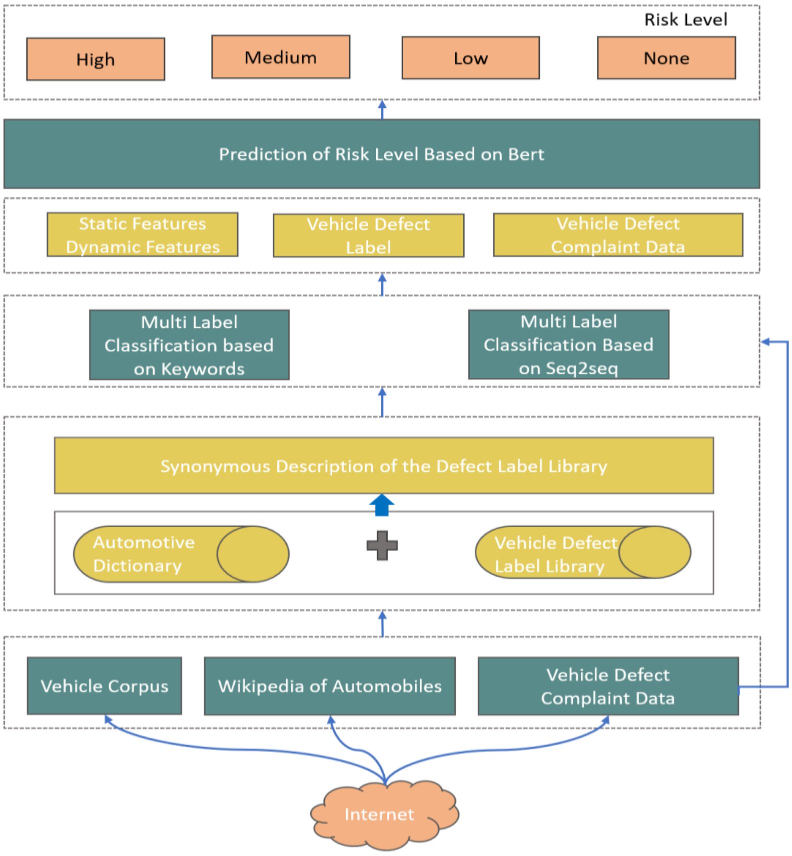

Fig. 1. Architecture of VDRF Model

3.2. Defect Label Library Feature

The vehicle defect label library consists of standardized vehicle defect names and corresponding typical defect descriptions. Embedding layer of the model used in this paper includes two parts, one part is based on the word vector. And the other part reflects whether the key words in the defect description appear corresponding vehicle defect description directly. Considering that the complaint data are from different kinds of consumers of different cultural levels, different descriptions may appear for the same group of different users of the defect, we expanded the synonym of the existing defect label library in this part. After analysis, the defect description is usually composed of secondary assembly and specific defect description, such as "door rust". The secondary assembly is mainly the name of the vehicle parts. We extend the nickname, abbreviation and common misnomer of vehicle parts by search engine. For the vehicle defect description part, we use the synonym extension tool synonyms [22] to extend this collection. We replace the word vector model of the toolkit with the pre-trained vehicle domain word vector. Candidate words are selected by similarity of defect description.

Finally, a defect label library with extended synonymous descriptions is obtained. In the embedding layer of the model, the representation of a word is divided into two parts, one is the word vector represented by the domain word vector model, and the other is the 32-bit defect coding feature bits trans-formed from the defect coding. For each word in the complaint text, if the current word belongs to the defect label library or the corresponding secondary assembly appears in the text, the word defect coding feature position of the complaint text is defect code, otherwise the defect coding bit of the word is ‘0000’.

Table 1. Vehicle defect label library code

First assembly	Second Assembly	Defect Label	Defect Code
车身 car body	车门 doors	车门生锈 Rusting of doors	5002
车身 car body	车门 doors	车门缝隙 doors gap	5007
发动机 engine	进排气系统 Intake and exhaust	排气管脱落 pipes fall off	2104
发动机 engine	点火与起动系统 starting system	喷油嘴故障 Injector fault	2205
制动系统 brake	制动通用装置 brake device	回位不良 return fault	6310

3.3. Multi-label Classification of Vehicle Defect Information Collection based on Seq2seq Model

The basic idea of seq2seq is using Bi-LSTM as encoder to read the input sentence, that is, the whole sentence is compressed into a fixed dimension of the code, and then use another LSTM called decoder to read the code, the information of the sentence will be compressed into a vector. And the architecture of the multi-label classification of vehicle defect information is shown in Figure 2.

Embedding. Firstly,Word segmentation tool jieba [23] with the vehicle domain dictionary constructed in our previous published paper [24] is employed on the complaint text S. Then, the segmented complaint text S is vectorized in the embedding layer, which can reduce the input dimension and reduce the number of parameters of the neural network. Furthermore, the dense vector representation of the word vector layer can contain more semantic information [25].

Encoder layer. Bidirectional LSTM [26](Bi-LSTM) recurrent neural network is used to read the text information in order from the front and back two directions, and to calculate the hidden layer vector h_i for each word w in the complaint text S. Each word corresponds to the hidden state vector h, which includes the state vectors in the two directions $\vec{h}_i$. And $\overleftarrow{h}_i$ representing the semantic information centered on i^{th} word.

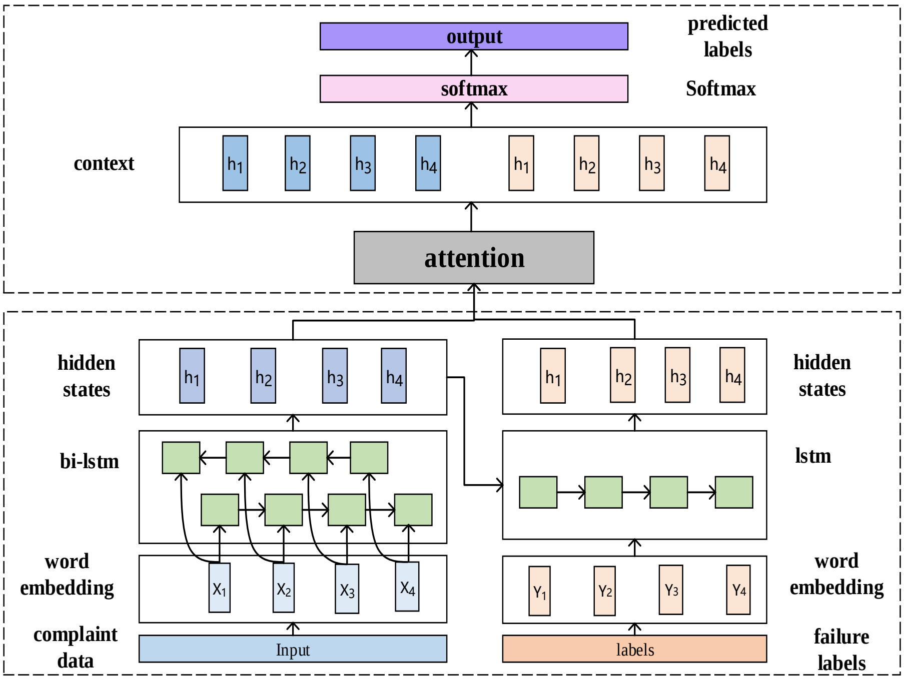

Fig. 2. Architecture of Multi-label Classification of Vehicle Defect Information

Attention mechanism. Due to the different words have different effects on prediction labels, seq2seq model with attention mechanism is used to find out the hidden state of encoder and decoder through attention connection.

Decoder layer. LSTM recurrent neural network is used in decoder layer. The decoder receives the hidden layer state s_{t-1} at time-step t, the context vector $c_{t-1}$and the label distribution vector $l(y_{t-1})$ from the attention mechanism, respectively, and inputs them to the decoder. The vector $l(y_{t-1})$ reflects the overall distribution of labels. Vector $l(y_{t-1})$ is added to the decoding process can integrate the relationship between labels.

Softmax layer. Softmax is used as in the classification lyaer, and a defect label y_t with the highest probability is generated by the output state vector s_t from the decoder.

3.4. Vehicle Defect Recall Risk Rating Forecast Model based on Bert Model

Automobile complaint information derived from the defect information collection system of the Defective Product Management Center is used as the data source, and the automobile defect risk recall risk data set contains about 120,000 pieces of defect information. After sorting out these pieces of defect information, we finally form 10,351 typical automobile defect cases.

By analyzing the defect information data source, the main composition of defect information is shown as in Table 2.

Table 2. Defect information composition

Item number	Item	Data sample
1	缺陷信息编号 Defect information number	QC201312001
2	时间 Time	201312
3	缺陷信息来源 Defect information source	备案 Put on record
4	生产者 Producer	阿斯顿马丁拉共达（中国）汽车销售有限公司 Aston Martin Lagonda (China) Automobile Distribution Co. Ltd
5	品牌 Brand	Aston Martin(阿斯顿马丁) Aston Martin
6	车型 Vehicle type	V8 WANTAGE
7	里程信息 Mileage information	1.38 万公里 13800 km
8	使用年限 Service life	2.00 年 2 years
9	总成 Assembly	发动机 Engine
10	分总成 Sub assembly	汽油发动机 Petrol engine
11	故障标签 Defect label	离合器液压软管夹失效 Clutch hydraulic hose clamp failure
12	缺陷描述 Defect description	离合器液压软管夹失效可能导致油泄漏和离合器失效，需更换新型管夹 Failure of clutch hydraulic hose clamp may cause oil leakage and clutch failure, so new clamp shall be replaced
13	故障等级 Fault level	中 Medium
14	舆情信息影响力 Influence of public opinion information	0
15	投诉数量	10

	Number of complaints	
16	召回风险（信息会商结果） Recall risk	低 Low

Defect information is mainly composed of three parts. The first part is the basic information of automobile, such as manufacturer, brand, model, mileage information and so on. The second part is the defect information of automobile fault, including the assembly and sub-assembly, and the third chapter is the result of multi-label classification of complaint information based on the defect management center automobile fault classification system, in addition, there is a defect fault severity evaluation level.

Public opinion information refers to the daily monitoring of public opinion information based on the quality and safety of automotive products and special monitoring of public opinion for specific events, mainly including public network media and We Media two sources.

In the process of defect recall, the result of information consultation is a pre-liminary judgment of recall risk made by the staff of automobile recall management according to the typical complaint information in a period of time. There are four levels, "high" means higher recall risk, "medium" means general recall risk, "low" means lower recall risk, and "none" means almost no recall risk. In-formation conferences are held quarterly to discuss the risk levels of some typical defect cases and to take different recall management measures for cases with different risk levels. This paper selects the results of the first information consultation as the correct risk level of the case data. In the following comparative experiments, the results of the latest meeting were selected to do the corresponding comparison and analysis.

In order to facilitate the subsequent processing of automobile defect information, it is necessary to preprocess the data in the data set. First remove duplicate and similar defect information, change the null and missing values to default values, and then normalize the car brand, manufacturer, and model. We find that there are some ambiguities in the fault labels, the fault labels are standardized in the defect information according to the classification system of the defect management center.

Automobile defective product risk recall prediction is actually a multi-classification problem. It can be found from the table that the amount of data of different risk levels is quite different, therefore, it is very important to solve the imbalance problem of data category during multi-classifying.

The purpose of data analysis is to extract the key information which may reflect the risk of automobile recall from the above defect information. After many discussions with experts of automobile recall research in the Defect Management Center, we summarized three kinds of characteristics: static risk characteristics, dynamic semantic characteristics and fault semantic characteristics. Static characteristics mainly include the brand, model, manufacturer, and defect information types. Dynamic features include mileage and years of car purchase. Fault semantic features include fault labels and fault severity levels. These features can describe and reflect the risk information hidden in defect information from different dimensions. The static features can be obtained directly from the dataset, while the dynamic features can be obtained from the defect information of the latest time stamp. Fault label features are selected from the assembly and fault label through natural language processing techniques. These features will be used as the input of the model, and provide a comprehensive and rich feature basis for the automobile defective product recall risk prediction model.

Through the statistics and analysis of the data set of automobile defect cases, we find that the automobile fault description and the automobile fault label also have certain influence on the recall risk level. The SVM model mentioned above only deals with numeric features, which is unable to capture semantic information in the fault description. The semantic representation of the text directly determines the accuracy of vehicle recall risk prediction. Bert language pre-training model is firstly used in this paper to predict the risk of automobile recall.

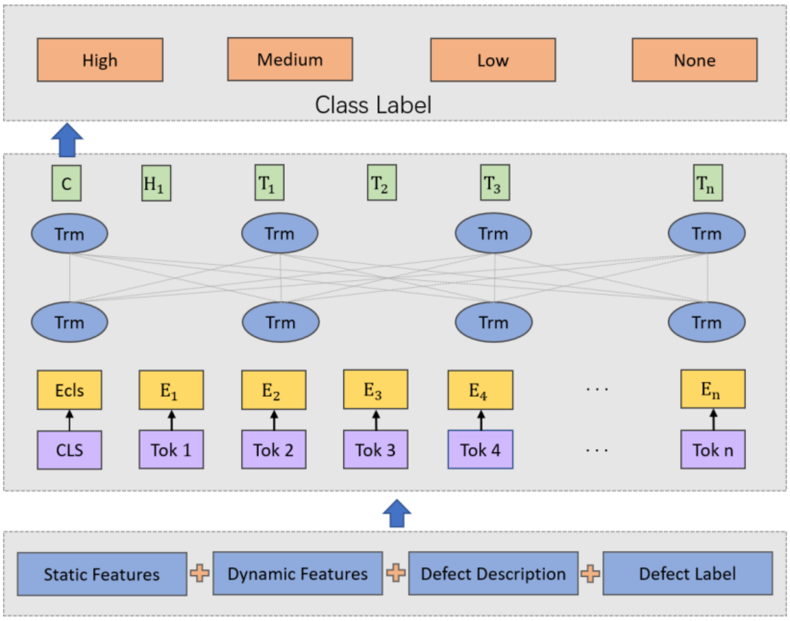

Fig. 3. The Architecture of VDRF based on Bert Model

The prediction task of automobile recall risk level in this paper can be regarded as a basic text classification task. Therefore, the modification of the network structure is very simple, only the first output of the last layer of Transformer needs to be used as the sentence label.

The model structure diagram of the modified Bert model used in VDRF is shown in Figure 3.

The core idea of attention mechanism used by Transformer is to calculate the relationship between each word in a sentence and all the words in the sentence, and then to think that the relationship between these words reflects the relevance and importance of different words in the sentence to some extent. Therefore, by using these relationships to adjust the importance (weight) of each word, a new expression of each word can be obtained. This new representation not only contains the word itself, but also contains the relationship between other words and the word, so it is a more global expression than a simple word vector. Transformer obtains the final text representation by continually overlapping the input text with this attention mechanism layer and the normal non-linear layer.

Therefore, the prediction of recall risk is essentially a supervised multi-classification problem, in order to accurately predict the risk level from the multi-dimensional heterogeneous defect information characteristics. In this paper, we use SVM and Bert models to predict the risk of different feature combinations of defect cases based on the existing machine learning and deep learning technologies.

Because SVM is suitable for dealing with discrete data, it is necessary to deal with the static, dynamic and fault characteristics first. For numeric class features, they are entered directly into SVM, and for class features, the one-hot method is used to convert them to 0-1 vectors. For models, assembly information, and fault labels, an index dictionary is built to numeralize the features. And then normalize the features. By selecting the appropriate kernel function and decision function, a text classifier can be obtained.

Because SVM ignores the semantic information in the fault features, natural language processing technology is used to obtain the semantic information in this paper, which may reflect the recall risk of the vehicle from another perspective.

Static features, dynamic features and fault features are used as sequence input data. Then, based on the pre-trained Chinese Bert model, the last layer of the network is reconstructed, and the text classification task based on Bert is done.

4. Experimental Results and Analysis

In this section, we evaluate method of the multi-label classification of vehicle complaint data and the method of the risk level prediction in the corresponding corpus. The corpus used in the experiments will be described firstly. Then the experimental results will be analyzed and discussed in the following sections.

4.1. Experimental Datasets

DPAC Corpus. This dataset is provided by the defect information collection system of Defective Product Administrative Center. It contains more than 130,000 pieces of vehicle defect complaint information, which contain one or more defect labels marked by experts in 22,747 pieces of data. These defect labels are from the Vehicle Defect Label Library of the Defective Product Administrative Center, which contains 934 defect labels. The number of defect labels and the samples of data are listed in Table 3.

Table 3. DPAC corpus Statistical tables

The number of label	1	2	3	>=4
22747	16351	4991	1183	222
Percentage	71%	23%	5%	1%

AUTO Corpus. It is a new large dataset form a vehicle complain website by our crawler system. It contains more than 200,000 descriptions of complaints about defects in vehicles. All of the defect information is labeled by experts. These defect labels come from the vehicle defect classification label library of the vehicle complain website, with a total of 402 defect labels. The number of defect labels and the samples of data are listed in Table 4.

Table 4. AUTO corpus Statistical tables

The number of label	1	2	3	>=4
200000	136701	44814	12871	5601
Percentage	68%	22%	6%	4%

DCRL Corpus. The automobile defect case risk level data set contains 10,351 typical automobile defect cases. The statistics of different risk levels are shown in Table 5 below.

Table 5. DCRL corpus Statistical tables

Risk level	High	Medium	Low	None	Total
Number	854	704	2944	5849	10351
Percentage	8.4%	6.7%	28.4%	56.5%	100%

4.2. Evaluation Metrics

Hamming-loss [27], Micro-F1 [28] and Macro-averaging are used indicators in multi-label classification tasks [29].

4.3. Experimental Details

Our experiments have two main parts. The first is multi-label classification experiments. And the other is risk prediction experiments. For multi-label classification experiments, the most representative multi-label classification algorithms are selected as baseline, and the comparative experiments are carried out in large-scale corpora (AUTO corpus) and small-scale corpora (DPAC corpus).

In Multi label classification experiment, the pre-trained vehicle domain word vector model is used as word representation. In order to avoid the impact of the vehicle brand on the prediction result, synonymous substitution of the description of the vehicle brand and the vehicle system is used, and the corresponding substitution of the figures in the complaint text are also used. After statistical analysis, the first 600 words of the complaint text are intercepted as input, and the part exceeding the length of the complaint text will be discarded. Referring to the conclusion of paper [14], the frequency of the defect labels corresponding to the complaint text in the training data is sorted. The hidden state vector of the encoder and decoder is set to 300 and 600 respectively, and the number of LSTM layers of the encoder and decoder is set to 2. In the training phase, the loss function is the cross-entropy loss function. Adam optimizer is used to minimize the cross-entropy loss function [15].

That is the detailed information set during the experimenting is shown as in Table 6

Table 6. Parameters setting in the experiments

Parameters	Value
Word embedding dimension	200
Label feature dimension	32
Length of beam search	5
Number of hidden layers in encoder	300
Number of hidden layers in de coder	600
Learning rate	0.001
Dropout of Learning rate	0.5
Optimizer	Adam
epoches	2000

For the experiment of risk prediction, two methods are employed, one is based on SVM model, the other is based on Bert model. The automobile defect case data set are divided into two groups. For each risk category, 80% are selected as the training data and 20% as the test data. At the same time, different features were selected to carry out multiple sets of contrast experiments to predict the risk level of defective vehicle recall. In this experiment, there are three kinds of features, which are static features, dynamic features and fault features obtained from the automobile defect label classification experiment. Combination of these three characteristics is used as the input of the recall risk prediction model, and comparative experiments are carried out.

4.4. Experimental Results and Analysis

Multi Label Classification Model

In order to evaluate the performance of different multi-label classification methods, the following five representative methods are implemented on the two dataset.

Binary Relevance (BR) [3]: transforms each label in multiple labels into a single label classification problem.

Classifier Chains (CC) [5]: transforms the multi-label classification problem into a single label classification problem, which introduces the relational information between labels in a chain structure of one label.

Label Powerset (LP) [6]: treats every possible label set combination as a new label, transforming the problem into a multi-classification problem with a single label.

CNN-RNN [12]: Global and local text semantics and label dependencies are captured using CNN and RNN, and label sequences are predicted using RNN.

The Sequence Generation Model (SGM) [14]: transforms the multi-label classification problem into a sequence generation problem, and generates a label sequence using a global-embedding decoder architecture.

We implement the BR and CC algorithms using the open source multi-label classification toolkit Scikit-Multilearn [31], and use Support Vector Machine (SVM) as the basic classifier in these algorithms [32][33].

Table 7. Label prediction results comparison

Corpus	AUTO		DPAC	
Metrics	Hamming Loss	Micro-F1	Hamming Loss	Micro-F1
BR-BF	0.0106	0.5996	0.0529	0.5517
BR-W2V	0.0038	0.6301	0.0319	0.6103
CC-BF	0.0087	0.6176	0.0473	0.5885
CC- W2V	0.0031	0.6565	0.0297	0.6237
LP-BF	0.0097	0.6028	0.0476	0.5904
LP-W2V	0.0032	0.6468	0.0415	0.6175
CNN-RNN	0.0031	0.6971	0.0178	0.6412
SGM	0.0027	0.7203	0.0125	0.6563
Seq2seq	0.0028	0.7195	0.0129	0.6511

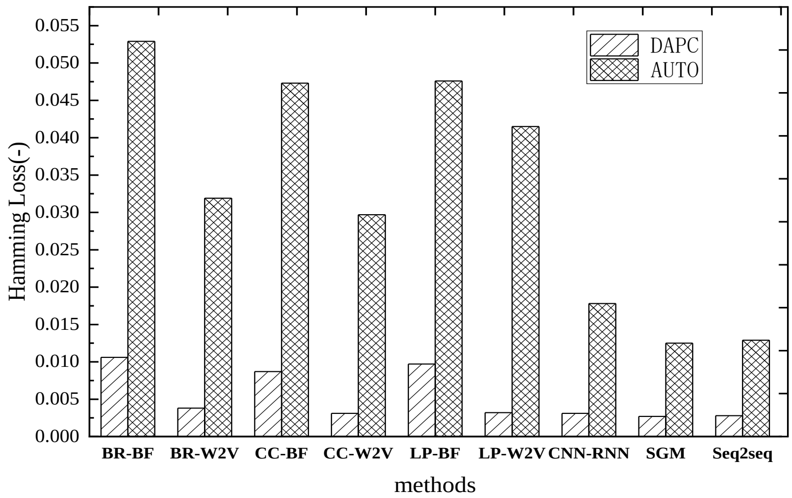

Fig. 4. Comparison of Hamming Loss

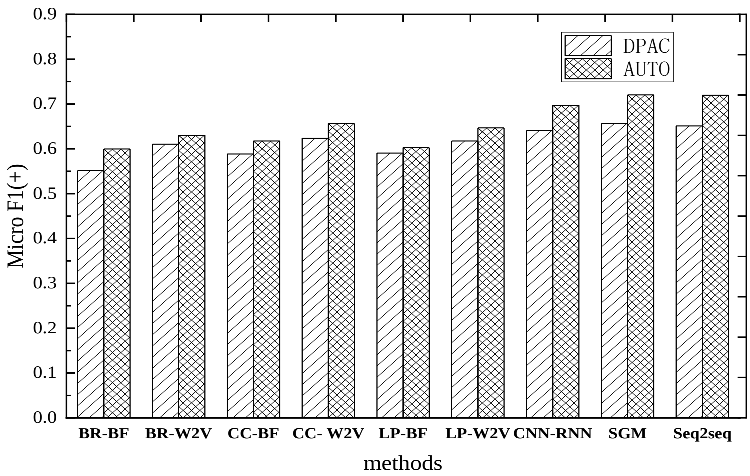

Fig. 5. Comparison of Micro-F1

Based on pre-trained vehicle domain word vectors, five typical multi-label classification methods are tested on two vehicle complaint datasets. The experimental results are shown in the following Table 7, Figure 4 and Figure 5, where BR stands for Binary Relevance algorithm, CC stands for Classifier Chains algorithm, BF stands for feature extraction based on vehicle defect labels, and LE stands for adding defect labels distribution vectors at the decoding layer.

In BR, CC, and LP algorithms, for a complaint text containing m words, the pre-trained domain word vector model is used to obtain the word representation vector of each word, and then the average value is obtained to represent the complaint text.

The following conclusions can be drawn from the above experiment results:

(1) Neural network-based methods are better than those using traditional multi-label classification, which shows that the neural network can recognize text information better and improve the accuracy of classification in multi-label classification.

(2) In the traditional machine learning multi-label classification method, the selection of text features has a great influence on the prediction results. From the table, it can be seen that for the same method, the result of using pre-trained domain word vectors is better than that of using label-only database features to express the complaint text, which verifies the necessity of pre-trained domain word vector model.

(3) Compared with the BR algorithm and the CC algorithm, the Classifier Chains algorithm performs better because the multiple defect descriptions contained in the vehicle complaint data are generally related to each other, and the CC algorithm takes into account the relationship between the labels. Because LP algorithm transforms the problem of multi-label classification into the problem of multi-class classification in single-label learning, and there are many kinds of multi-label combinations in the data analysis and statistics, LP algorithm is not suitable to solve this problem, and the experimental results also prove this point.

(4) Compared with CNN-RNN model, seq2seq model performs better in multi-classification of Chinese complaint texts. The reason is that seq2seq model reads the semantic information before and after each word in the complaint texts through Bi-LSTM, and pays attention to the words related to the predicted failure results through attention mechanism. CNN-RNN focuses on the high-order relevance of labels, but the recognition of the semantic information of the text itself is insufficient.

(5) Comparing SGM model with seq2seq model with attention mechanism, the input of SGM model and seq2seq model is based on pre-trained vehicle domain word vector model, and the value of word vector is allowed to change during the training process, because SGM model is based on seq2seq model with mask module and global embedded information (global embedded) in the decoder part. Experiments show that the mask module and global embedding vector are equally effective in vehicle complaint dataset. In analyzing the classification results of seq2seq model, we also find that the prediction results of the same article text contain some duplicate labels.

Based on the above conclusions, we add the feature of extended vehicle defect label library (CF) to the input layer of seq2seq model with attention mechanism. Considering the diversity of vehicle defect label combinations, a label distribution vector (LE) of each vector is obtained by using the training method of word2vec based on the defect label text of all data. A comparative experiment was carried out in two datasets. The results are shown in Table 8, Figure 6 and Figure 7.

Table 8. Label prediction results comparison

Corpus	AUTO		DPAC	
Metrics	Hamming Loss	Micro-F1	Hamming Loss	Micro-F1
Seq2seq	0.0028	0.7195	0.0129	0.6511
SGM	0.0027	0.7203	0.0125	0.6563
Seq2seq+CF	0.0026	0.7212	0.0121	0.6532
Seq2seq+CF+LE (VDIF-M)	0.0025	0.7363	0.0100	0.6624

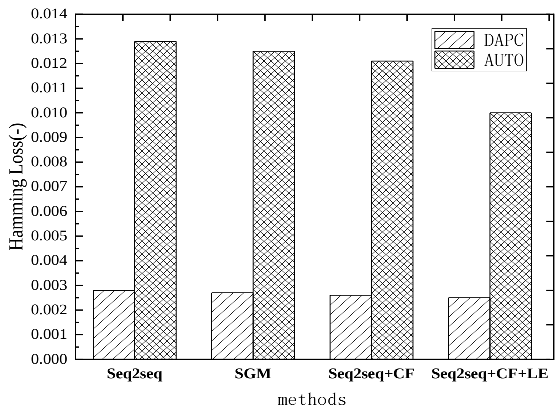

Fig. 6. Comparison of Hamming Loss

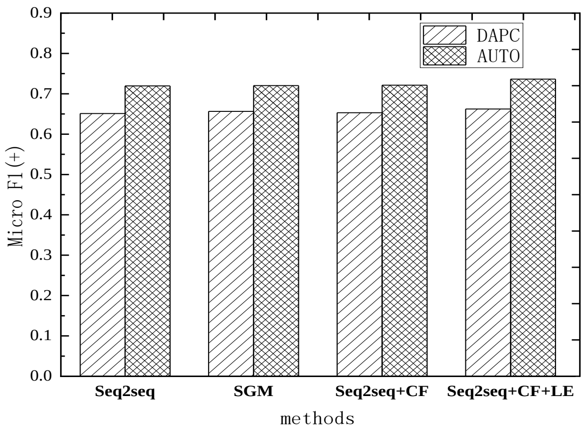

Fig. 7. Comparison of Micro-F1

The experimental results in the table show that the label library features added have obvious effect on the auto dataset, and the reason may be that there are fewer defect categories in the vehicle quality network, but there are more defect labels in the dataset of DPAC corpus, so the effect of adding label library features is not obvious. After the label distribution vector is added to the decoder layer, it is improved both in two datasets. Comparing with the SGM model, the experimental results show that the proposed method is superior to the SGM model in two datasets, because our methods adds defect label features suitable for vehicle complaint data, and uses the pre-trained domain word vector model at the same time.

Table 9 shows some instances of a multi-label classification that uses the different sequence models to identify only the “Engine Abnormal Noise” label in the defect description. Our proposed VDIF-M model can not only recognize the "engine-abnormal noise" label, but also generate the "Body Vibration" label ac-cording to those words "vehicle" and "jitter". This is because the extended fault description synonymous label library contains synonymous relationships between "vehicle resonance" and "vehicle jitter", which verify the model proposed in this paper can solve the multi-label classification problem of some instances by adding defect label features.

Table 9. DCRL corpus Statistical tables

Defect description	VDIF-M	Seq2seq	Correct Label
发动机有明显异响，我不懂车都能听出来，而且车辆抖动，去店里检查，说什么都正常，抖动也正常。	发动机-异响 车身附件及电器-车身共振	发动机-异响	发动机-异响 车身附件及电器-车身共振
The engine is obviously abnormal,	Abnormal	Abnormal	Abnormal engine

don't understand the car can hear, and the car jitter, go to the store to check, say what is normal, jitter is normal.	engine noise Body Vibration	engine noise	noise Body Vibration
挂d挡速度上升到40时发动机转速达到4000，但车速不上升；挂 r 挡后退无力踩住刹车时，车身抖动严重．去4s店检测,说是变速箱的3-5模块损坏,要大修变速箱。	发动机-无法提速 变速器-电脑板故障	发动机-无法提速 变速器-异响	发动机-无法提速 变速器-电脑板故障
When the speed of the gearbox increases to 40, the speed of the engine reaches 4000, but the speed of the car does not rise; when the gearbox is unable to step on the brake, the body shakes seriously. Go to 4S shop to check that the 3-5 module of the gearbox is damaged, it is necessary to overhaul the gearbox.	Engine Unable to Speed up Transmission-Computer Board Failure	Engine Unable to Speed up Transmission Abnormal engine noise	Engine Unable to Speed up Transmission Computer Board Failure

Recall Risk Prediction Model

In order to accurately predict the risk level from the multi-dimensional heterogeneous defect information features, SVM and Bert models are used to predict the risk of different defect case feature combinations based on the existing machine learning and deep learning technologies

In the SVM experiment, one-hot to represent the class information in the static feature directly. For the automobile brand, manufacturer and other information, an index table is built to convert the corresponding features into numerical values. The mileage in the dynamic features is in the unit of 10,000 km and the service life is in the unit of years. In this experiment, we first use different feature combinations, choose the kernel function as Gaussian kernel, and the penalty coefficient is 1, the class weight is the default.

From Table 10, it can be seen that different combinations of features have different effects on recall risk. Static features have the greatest impact on recall risk, dynamic features have the least impact, and the effect of three types of features fusion is the best.

Table 10. Different Feature Combination Result-SVM

Feature combination	Parameters	Macro-acc	Macro-recall	Macro-f1
$\vec{X}s$	default	0.62	0.55	0.57
$\vec{X}d$	default	0.55	0.56	0.55
$\vec{X}l$	default	0.60	0.54	0.56
$\vec{X}s\,\vec{X}d$	default	0.62	0.67	0.64
$\vec{X}s\,\vec{X}l$	default	0.65	0.69	0.66
$\vec{X}d\,\vec{X}l$	default	0.61	0.55	0.57
$\vec{X}s\,\vec{X}d\,\vec{X}l$	default	0.80	0.68	0.72

Although the overall results of the experiment are very good, for the categories with a small number of samples, the prediction ability of SVM is very limited.

In the experiment of Bert model [34], we forecast the recall risk of defect cases based on the pre-trained Chinese language model Bert model published by Google. Because Bert model is more suitable for processing sequence data, dynamic features, static features, defect features and the combination of the three features are used for comparative experiments. At the same time, we adjust the learning rate and the maximum sequence length in the training process to get the most suitable combination of parameters for this task. The achieved experimental results are shown in Table 11.

Table 11. Different Feature Combination Result-Bert

Feature combination	Macro-acc	Macro-recall	Macro-f1
$\vec{X}s$	0.65	0.59	0.61
$\vec{X}d$	0.59	0.56	0.57
$\vec{X}l$	0.61	0.59	0.6
$\vec{X}s\,\vec{X}d$	0.71	0.65	0.67
$\vec{X}s\,\vec{X}l$	0.73	0.62	0.67
$\vec{X}d\,\vec{X}l$	0.64	0.60	0.62
$\vec{X}s\,\vec{X}d\,\vec{X}l$	0.79	0.78	0.79

From Table 11, it can be found that the combination of the three features has achieved best results in the process of risk prediction, which is the most suitable parameter combination for this task.

Detailed metrics comparison between the Bert and SVM models are listed in Table 12. As can be seen from the above table, the prediction effect of Bert and SVM is not much different for the categories with more sample data. Bert model is more accurate in predicting the smaller sample categories, which also shows that Bert model can solve the sample imbalance problem to some extent.

Table 12. Detailed comparison between SVM and Bert

Model	Risk Level	Accuracy	Recall rate	F1-score	Number of samples
Bert	high	0.89	0.82	0.85	171
	medium	0.59	0.62	0.60	141

	low	0.80	0.80	0.80	589
	none	0.88	0.89	0.89	1170
	micro-avg	0.84	0.84	0.84	2071
	macro-avg	0.79	0.78	0.79	2071
	weighted-avg	0.84	0.84	0.84	2071
SVM	high	0.89	0.74	0.81	171
	medium	0.43	0.62	0.51	141
	low	0.90	0.84	0.87	589
	none	0.76	0.83	0.79	1170
	micro-avg	0.81	0.81	0.81	2071
	macro-avg	0.75	0.75	0.74	2071
	weighted-avg	0.83	0.81	0.82	2071

In most multi-classification tasks, the category with smaller sample size is usually the most concerned. In this task, "medium" risk and "high" risk category have the smaller sample size. Therefore, they are the most important supervision objects of defect recall management. In cases where the recall risk is neutral and high, appropriate measures are usually taken in the subsequent process. High-risk and medium-risk samples are similar, and the characteristics are not clear in the process of multi-classification, which is also one of the reasons for the low accuracy of high-risk and medium-risk cases in this topic.

5. Conclusion and Future Work

In the management of automobile recall, the risk assessment of automobile defect cases is the basis of the follow-up supervision. In order to improve the efficiency of defect risk rating assessment, this paper presents a model for predicting the risk level of automobile recall based on defect information. After building the defect case dataset, the multi-dimensional features of case data are extracted by data analysis, Bert model is used to read the fault information from the defect information, and different feature combinations and different models are compared with each other. The experiments show that the recall risk prediction model based on Bert model has best performace both in classification and prediction tasks. which can provide a powerful reference for automobile defect experts.

In the research of automobile recall risk prediction, only the defect information of automobile recall cases is collected by the defective product management center, most of which are the inherent attribute values and dynamic attribute characteristics of automobile. Although it also contains a public opinion index feature, the public opinion features of each case are not different. In the follow-up risk prediction work, mining the relevant industry news, complaint news, user comments and other information of different automobile brands is the main work in the future of automobile recall risk level prediction.

Acknowledgements. This work is supported by National Natural Science Foundation of China under Grants No. 61671070, 61972364. Defense-related Science and Technology Key Lab Fund project 6142006190301, National Language Committee of China under Grants ZDI135-53, and Project of Developing University Intension for Improving the Level of Scientific Research–No.2019KYNH226, Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University No.QXTCP B201908.

References

1. Defective Product Administrative Center Homepage, http://www.dpac.gov.cn, last accessed 2019/1/24.
2. Boutell, M. R., Luo, J., Shen, X., Brown, C. M.:Learning multi-label scene classification. Pattern recognition.37, 1757-1771.(2004)
3. Zhang, M. L., Wu, L.:Lift: Multi-label learning with label-specific features. IEEE transactions on pattern analysis and machine intelligence, 37, 107-120.(2014)
4. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3, 1-13.(2007)
5. Read, J., Pfahringer, B., Holmes, G., Frank, E.:Classifier chains for multi-label classification. Machine learning, 85, 333.(2011)
6. Clare, A., King, R. D.:Knowledge discovery in multi-label phenotype data. In European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg,42-53.(2001)
7. Gao, H.H., Liu,C., Li, Y.H.Z., Yang, X.X. V2VR: Reliable Hybrid-Network-Oriented V2V Data Transmission and Routing Considering RSUs and Connectivity Probability[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, DOI: 10.1109/TITS.2020.2983835
8. Elisseeff, A., Weston, J.:A kernel method for multi-labelled classification. In Advances in neural information processing systems.681-687.(2002)
9. Zhang, M.L., Zhou, Z.H.:ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 40,2038-2048.(2007)
10. Li, L., Wang, H., Sun, X.:Multi-label Text Categorization with Joint Learning Predictions-as-Features Method[C]// Conference on Empirical Methods in Natural Language Processing.835-839.(2015)
11. Zhang, M. L., Zhou, Z. H.:Multi-label Neural Networks with Applications to Functional Genomics and Text Categorization [J]. IEEE Transactions on Knowledge & Data Engineering, 18,1338-1351.(2006)
12. Chen, G., Ye, D., Xing, Z.:Ensemble application of convolutional and recurrent neural networks for multilabel text categorization[C]// International Joint Conference on Neural Networks. IEEE, 2377-2383.(2017)
13. Baker, S., Korhonen, A.:Initializing neural networks for hierarchical multi-label text classification. InBioNLP.307-315.(2017)
14. Yang, P., Sun, X., Li, W.:SGM: Sequence Generation Model for Multi-label Classification [J].(2018)
15. Gao, H.H., Kuang, L., Yin, Y.Y., Bin Guo, B.,Dou, K.Mining Consuming Behaviors with Temporal Evolution for Personalized Recommendation in Mobile Marketing Apps [J]. ACM/Springer Mobile Networks and Applications (MONET), 2020, DOI: 10.1007/s11036-020-01535-1
16. Zhang, X., Niu, S., Zhang, D.; Wang, G. A.; Fan, W.:Predicting vehicle recalls with user-generated contents: A text mining approach. In Pacific-Asia Workshop on Intelligence and Security Informatics. Springer, Cham.41-50.(2015)
17. Yang, S.L. :Research on Early Warning Model of Defective Automobile Recall Risk Based on Internet Complaint Data [D].Beijing Information Science and Technology School.(2015)

18. Jiang, C. Q, Wang, Q.L., Liu, S.X., Ding, Y., Liu, R. :Semi supervised learning method for vehicle defect recognition in Chinese social media environment [J]. Chinese Journal of Management Science, 22,677-685.(2014)
19. Abrahams, A. S., Jiao, J., Wang, G. A., Fan, W.:Vehicle defect discovery from social media. Decision Support Systems.54, 87-97.(2012)
20. Abrahams, S., Jiao, J., Fan, W., Wang, G. A., Zhang, Z. :What's buzzing in the blizzard of buzz. Automotive component isolation in social media postings Decis. Support Syst, 55.(2013)
21. Gao, H.H., Xu, Y.S.,Yin, Y.Y., Zhang, W.P., et al. Context-aware QoS Prediction with Neural Collaborative Filtering for Internet-of-Things Services [J]. IEEE Internet of Things Journal(IoT-J) , 7(5): 4532-4542 (2020)
22. Chinese synonyms Toolkit. https://github.com/huyingxi/Synonyms, last accessed 2019/1/24.
23. Chinese Word Segmentation Tool. https://pypi.org/project/jieba/, last accessed 2019/1/24.
24. Zhang, Y.W., Li, B.A., Lv, X.Q. :Research on Domain Term Dictionary Construction based on Chinese Wikipedia, Image Processing, Computing and Big Data .(2018)
25. Mikolov, T., Chen, K., Corrado, G., Dean, J. :Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.(2013)
26. Graves, A., Schmidhuber, J.:Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18, 602-610.(2005)
27. Schapire, R. E., Singer, Y.:Improved boosting algorithms using confidence-rated predictions. Machine learning, 37, 297-336.(1999)
28. Manning, C. D., PRABHAKAR, R., HINRICH, S.:Introduction to information retrieval, volume 1 Cambridge University Press. Cambridge, UK.(2008)
29. Ma, X.J., Gao, H.H., Xu, H.H., Bian, M.J.. An IoT-based task scheduling optimization scheme considering the deadline and cost-aware scientific workflow for cloud computing. EURASIP Journal on Wireless Communications and Networking [J], 2019, 2019(249). (2019)
30. Wiseman, S., Rush, A. M.:Sequence-to-sequence learning as beam-search optimization. arXiv preprint arXiv:1606.02960.(2016)
31. Szymański, P., Kajdanowicz, T. A. :scikit-based Python environment for performing multi-label classification. arXiv preprint arXiv:1702.01460.(2017)
32. Yang, X.X., Xiaoxian, Zhou,S.J., Cao, M. . An Approach to Alleviate the Sparsity Problem of Hybrid Collaborative Filtering Based Recommendations: The Product-Attribute Perspective from User Reviews. MOBILE NETWORKS & APPLICATIONS, 25(2):376-390 (2020)
33. Zhu, R., Yu, D., Ji, S., Lu, M. :Matching RGB and Infrared Remote Sensing Images with Densely-Connected Convolutional Neural Networks. Remote Sens.11, 2836.(2019)
34. Lin, C., Miller, T., Dligach, D., Bethard, S., Savova, G. :A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In Proceedings of the 2nd Clinical Natural Language Processing Workshop.65-71.(2019)

Xindong You is currently an associate professor of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research at Beijing Information Science & Technology University, China. She holds a post-doctoral position at Tsinghua University with the Beijing Institute of Graphic Communication, from 2016 to 2018. Before as a post-doctoral, she is an Associate Professor in Hangzhou Dianzi University. Before joining Hangzhou Dianzi University, she was a Ph.D. student with Northeastern University, from 2002 to 2007. She received her PhD degree in 2007. She is in charge of the National Nature Science Foundation of China from 2014 to 2017 and Nature Science Funding of Zhejiang Province from 2013 to 2015. She has authored about 30 papers in the international conference or journals, most of them are indexed by EI or SCI database. Her current research areas include Natural Language Processing, Image Processing, Distributed Computing, Cloud Storage, Energy Management, Data Replica Management, etc.

Jiangwei Ma is currently a postgraduate student in Beijing Information Science & Technology University, China. She received her B.S. degree from Beijing Information Science & Technology University, China, in 2018. Her current research areas include Relation Extraction, Natural Language Processing, etc.

Yuwen Zhang was a graduate student at Beijing Information Science & Technology University, China. He received his master degree from Beijing Information Science & Technology University, China, in 2019. His current research areas include Automobile Risk Level Detection, Natural Language Processing, etc.

Xueqiang Lv is a professor in Beijing Information Science & Technology University. Before joining in Beijing Information & Technology University, he is a post-doctoral of Peking University from 2003 to 2005. Before as a post-doctoral, he is a PhD candidate in Northeastern University from 1998 to 2003. He received his PhD degree in 2003. Until now, he has been in charge of the National Nature Science Foundation of China three times. He has authored about 60 papers in the international conference or journals, most of them are indexed by EI or SCI database. His current research areas include Cloud Computing, Distributed Computing, Natural Language Processing, Image Processing, Information retrieval, Machine Learning, Deep Learning, etc.

Junmei Han is currently an associate professor of Department of Systems General Design Institute of Systems Engineering. AMS. PLA, China. She received the M.S. degree in Military Communication and PhD degree in information and communication engineering from National University of Defense Technology, Changsha, China, in 2013 and 2019, respectively. Her current research areas include Complex System Theory, Natural Language Processing, Graph neural network, etc.

Received: September 03, 2020; Accepted: April 23, 2020