



VDIF-M: Multi-label Classification of Vehicle Defect Information Collection Based on Seq2seq Model

Xindong You¹, Yuwen Zhang¹, Baoan Li¹, Xueqiang Lv¹(✉),
and Junmei Han²

¹ Beijing Key Laboratory of Internet Culture and Digital Dissemination
Research, Beijing Information Science & Technology University, Beijing, China
lxq@bistu.edu.cn

² Laboratory of Complex Systems, Institute of Systems Engineering,
AMS, PLA, Beijing, China

Abstract. Classification and treatment of vehicle defect complaint data is an important link in the process of vehicle recall. Traditionally, the complaint data is classified by keyword matching method based on defect label library during the process of dealing with vehicle complaint data, which heavily relies heavily on the quality of the vehicle defect label library. The speed of traditional classification methods is rapid, but the accuracy is low. We transform the classification task of vehicle complaint data into a multi-label classification problem. Multi-label classification of vehicle defect information collection based on seq2seq model named VDIF-M is proposed in this paper. Firstly, a synonymous vehicle defect description label library is constructed based on the vehicle defect description data and vehicle domain corpus collected from various channels. Then a seq2seq model is proposed to solve the problem of multi-label classification of vehicle complaint data, which fuses the distribution relationship between labels. Substantial experimental results show that the proposed method outperforms previous methods in multi-label classification of vehicle complaint data.

Keywords: Multi-label classification · Seq2seq · Label generation · Deep learning

1 Introduction

With the continuous development of the vehicle industry, vehicles have become a necessity in people's lives. Data show that China is the largest country of vehicle production and sales country in the world. At the same time, the quality defects of vehicle products have also aroused people's concern and the complaints about vehicle quality defects appear on the Internet. In recent years, the recall system of defective vehicle products in China has been gradually improved. 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]. Most of these complaints contain one or more defect description information. The Defective Product Administrative Center needs to investigate and verify the vehicle

defects reflected in these complaints to determine whether to initiate a recall. But because different users have different understanding of the vehicle, the same kind of vehicle defect may be expressed in different ways, which brings great difficulties to the defective product management center for the analysis and processing of these complaints data. It is a feasible solution to classifying these complaints with multi-label using the multi-label classification technology in Artificial Intelligence and natural language processing and according to the corresponding defect classification system. In this paper, we transform the classification task of vehicle complaint data into a multi-label classification problem employed with seq2seq model.

Multi-label classification is an important problem in the field of natural language processing. Multi-label classification is a concept relative to single-label classification. Traditional single-label classification associates instance X with a single label L from a previously known finite set of labels L . A single label data set D consists of n instances $(x_1, L_1), (x_2, L_2), \dots, (x_n, L_n)$. The multi-label classification task associates a subset of labels S with each instance. Thus, the multi-label data set D is composed of n examples $(x_1, S_1), (x_2, S_2), \dots, (x_n, S_n)$. In a practical application scenario, an instance is usually associated with multiple labels in most cases. In this paper, a piece of vehicle complaint information may contain two or more kinds of vehicle defect information. Considering the great achievements of neural networks in natural language processing in recent years, we transform the multi-classification problem into a label generation method in this paper. Solving the multi-classification problem with sequence-to-sequence model (seq2seq) is popular in machine translation and generative text summarization. The seq2seq model used in this paper consists of an encoder and a decoder with attention mechanism. The encoder uses Bi-directional Long Short-Term Memory (Bi-LSTM) to read the semantic information of the vehicle complaint information on the one hand, and compares the complaint text with the vehicle defect description label library on the other hand, and extracts the defect description features. The decoder generates a label sequence through the LSTM based on the previously predicted label. Because different words in the complaint information contain different amount of defect information, the attention mechanism can distribute different weight to different parts. Therefore, this kind of neural network model can capture the feature of the complaint text better.

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. The constructed datasets contain descriptions of all kinds of complaints in the process of vehicle recall.*
- (2) *We firstly employ the seq2seq neural network model to solve the multi-label classification on vehicle complaint data. And 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) *Substantial experiments are conducted on the two constructed dataset with different deep learning models, the experiment results demonstrate that the proposed method outperforms current existing methods in multi-label classification of vehicle complaint dataset.*

The following sections are organized as follows. Section 2 introduces the relevant work. We describe our methods in the Sect. 3. In Sect. 4, we present the experiments and make analysis and discussion. Finally in Sect. 5 we conclude this paper and explore the future work.

2 Relate Work

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], which improves the classification effect by clustering the positive and negative instances to construct the characteristics of each label in the multi-label. Label Powerset (LP) [4] algorithm transforms the multi-label classification problem into a single-label multi-classification problem by treating each label set as a new independent label. The Classifier Chain (CC) algorithm [5] transforms the multi-classification tasks into a series of binary classification problems. The author combines the multi-labels into a sequence, and adds the predicted labels into the feature vector when predicting the new labels in the sequence, which can introduce the global information into the fusion of labels and the relationship between labels. However, CC algorithm is inefficient in solving the problem of more labels or more samples.

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. Elisseff [7] proposes Rank-SVM algorithm by introducing loss function to support vector machine (SVM). Zhang and Zhou [8] proposed an improved ML-KNN algorithm based on k-nearest neighbor algorithm to solve the multi-label classification problem. Li [9] 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.

With the successful application of deep learning in image and speech fields in recent years, some neural network models are also applied to multi-label learning tasks. Zhang and Zhou [10] 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 [11] uses a combination of CNN and RNN to represent the semantic information of the text and the higher-order features between the labels. Baker [12] will map to the rows of co-occurrence labels to initialize the final hidden layer of the CNN to improve the model effect. Yang [13] put forward that multi-label classification task should be regarded as sequence generation problem,

and used a new sequence generation model with a new decoder structure to solve the multi-label classification problem, and achieved good results.

3 VDIF-M: Multi-label Classification of Vehicle Defect Information Collection Based on Seq2seq Model

This section will introduce the details of the method used in this paper to solve the problem of multi-label classification of vehicle complaint data. An overview of the method used in this paper is given in Subject. 3.1. In Subject. 3.2, some preparatory work is described, including word vector training, data preprocessing and the extension of vehicle defect label library. Finally the details of seq2seq model structure used in this paper are present in Subject. 3.3.

3.1 Model Architecture of VDIF-M

In the task of multi-label classification, we use L to represent the defect label library corresponding to the vehicle complaint text, which contains h class defect labels. The task of multi-label classification is to generate a set Y of corresponding labels for each complaint text x containing n words. Y is a subset of the label library L , and Y contains one or more labels like L_n .

An overview of our proposed model is Fig. 1. Firstly, in the embedding layer, we use the pre-trained word vector v_i to join the coding vector b_i in the defect label library to form x_i as the input of the seq2seq model. In the encoder layer, we use bi-directional LSTM to read x_i to get the hidden layer state vector h , and combine the attention mechanism to get the context vector c_t at time t . The decoder layer receives these vectors and predicts the label distribution vector v_1 corresponding to the previous label, and then gets the distribution of the label sequence through softmax layer. According to the distribution, we can get the defect label sequence $L_1, L_2 \dots L_n$.

3.2 Defect Label Library Feature

The vehicle defect label library is composed of standardized vehicle defect names and corresponding typical defect descriptions. Since the embedding layer of the model used in this paper consists of two parts, one part is based on the word vector obtained by the pre-trained vehicle domain word vector model. And the other part reflects whether the key words in the defect description appear corresponding vehicle defect description directly. Considering that the complaint data come from different kinds of consumers of different cultural levels, different descriptions may appear for the same group of different users of the defect, we expands 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 [14] to extend this collection. We replace the word vector model of the toolkit with the pre-trained vehicle domain word

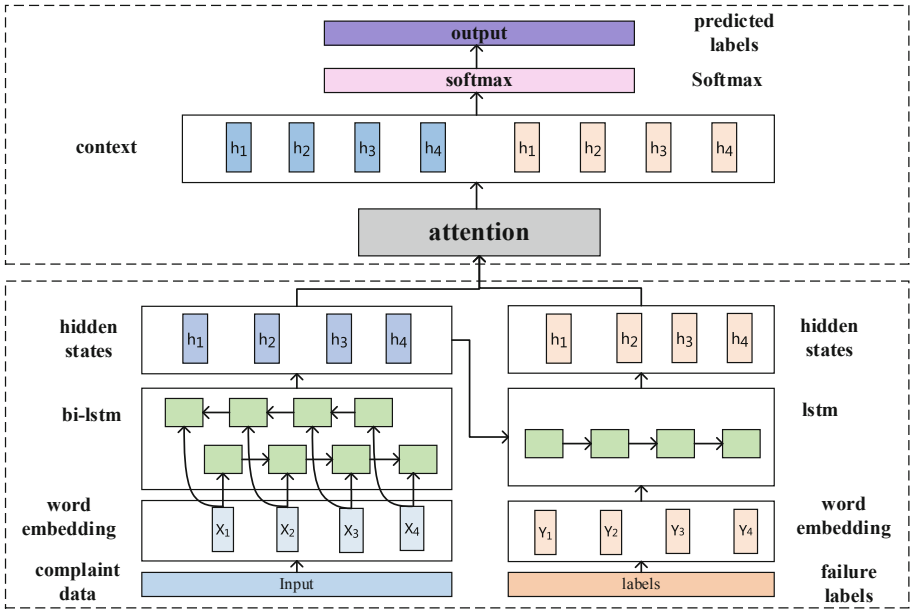


Fig. 1. VDIF-M model architecture

vector, that is to say, a group of synonyms is extended based on word2vec. Then the 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 transformed 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).

In order to capture the relationship between the defect labels corresponding to the vehicle complaint data, this paper first extracts the label data, each row corresponds to a set of defect labels of the complaint data, and converts the defect labels into codes according to the coding table of the vehicle defect label library. And then a vector v_1 which can reflect the distribution of defect labels is obtained by training word2vec word vector.

3.3 Seq2seq Model in Our Method

In this section, we introduce the seq2seq model used in this article in detail. The complete model includes the embedding layer, the encoder layer, the decoder layer, and the softmax layer. The basic idea of seq2seq is to use Bi-LSTM called encoder to read the input sentence, that is, the whole sentence is compressed into a fixed dimension of

Table 1. Vehicle defect label library code.

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

the code, and then use another LSTM called decoder to read the code, the information of the sentence will be compressed into a vector. This model is also called the encoder-decoder model.

Embedding. In a deep learning task, the quality of the word vector determines the final effect of the neural network. Embedding layer mainly vectorize the complaint text S . That is, the words in the text S are represented by a real vector, which can reduce the input dimension and reduce the number of parameters of the neural network. On the other hand, the dense vector representation of the word vector layer can contain more semantic information. After using the word segmentation tool jieba [15] for the complaint text S , a sequence of n words is formed and denoted by $w = (w_1, w_2 \dots w_n)$. In the process of word segmentation, in order to improve the accuracy of word segmentation, the vehicle domain dictionary constructed in our previous published paper [16] is used as the user-defined dictionary. The word2vec model proposed by Mikolov [17] is used to construct a pre-trained word vector model based on the vehicle domain corpus. The model forms $n*d$ embedding matrix, where n denotes the number of words in the dictionary and d denotes the dimension of the word vector. As described in the previous section, for the word w in complaint text, label feature vector \overrightarrow{wa}_i is constructed by searching whether the word in the text is the keyword of the vehicle defect label library. The purpose of this process is to capture label library features at the word embedding. The expression of the i_{th} word x_i in the complaint text is as follows:

$$\vec{x}_i = [\overrightarrow{wv}_i; \overrightarrow{wa}_i] \quad (1)$$

Where \overrightarrow{wv}_i is the word vector representation of the i_{th} word in the complaint text based on the pre-trained word2vec model, \vec{x}_i is composed of \overrightarrow{wv}_i and \overrightarrow{wa}_i splices.

Encoder Layer. In the encoder layer, we use a recurrent neural network bidirectional LSTM [18] to read the text information in order from the front and back two directions, and calculate the hidden layer vector h_i for each word w in the text. 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. The concealed state vector h is composed of the state vectors in the two directions. The calculation process is as follows:

$$\vec{h}_i = \overrightarrow{lstm}(\vec{h}_{i-1}, x_i) \quad (2)$$

$$\overleftarrow{h}_i = \overleftarrow{lstm}(\overleftarrow{h}_{i-1}, x_i) \quad (3)$$

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (4)$$

Attention Mechanism. When predicting defect labels, the complaint text may contain information that is not relevant to the defect label. Considering that different words have different effects on prediction labels, we use seq2seq model with attention mechanism to find out the hidden state of encoder and decoder through attention connection. The decoder searches the hidden state of encoder at every step of decoding by using the hidden state of encoder as the input of query calculating a weight related to the query input at each location of the input, according to this weight, the hidden state of each position is weighted to obtain a context vector. In decoding the next word, the context vector and the pre-trained label distribution vector label stitching are used as additional information input to the decoder, which enables the decoder to read the information most relevant to the vehicle defect in the text rather than relying entirely on the hidden vector at the previous moment. The attention mechanism assigns the vector context_i as follows:

$$e(h_i, s_j) = U_a \tanh(V_a h_i + W_a s_j) \quad (5)$$

$$\alpha_{i,j} = \frac{\exp(e(h_i, s_j))}{\sum_{k=1}^m \exp(e(h_i, s_k))} \quad (6)$$

$$\text{context}_i = \sum_{i=1}^m \alpha_{i,j} h_i \quad (7)$$

Where V_a, W_a, U_a are weight parameters and h_i is the hidden state.

Decoder Layer. In a decode layer, in order to capture the relationship between defect labels, we use the vector representation of the previous label based on the pre-trained label distribution vector and the context vector, and use the LSTM in the recurrent neural network. 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. Adding this vector to the decoding process can integrate the relationship between labels. The decoder calculates the hidden state vector s as follows:

$$s_t = LSTM(s_{t-1}, [l(y_{t-1}); c_{t-1}]) \quad (8)$$

Softmax. The softmax layer is the final prediction layer, and a defect label y_t with the highest probability is generated by the output state vector s_t from the decoder.

$$y_t = \left[\frac{\exp(V_l)}{\sum_{p=1}^L \exp(V_p)} \text{ for } l \text{ in } L \right] \quad (9)$$

Where L represents the vehicle defect label library and m_t is the mask vector. y_t is the label probability distribution at time-step t predicted by the model.

4 Experimental Results and Analysis

In this section, we evaluate our proposed methods on datasets. First, we introduced the datasets, evaluation metrics and experimental details. Then we make analysis and discussions about the experimental results.

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, of which about 22,747 pieces of data contain one or more defect labels marked by experts. 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 label and data sample is listed in Table 2.

Table 2. DPAC corpus Statistical tables

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

AUTO Corpus. We build a new large dataset form a vehicle complain website by crawler system. It contains more than 200,000 descriptions of complaints about defects in vehicles. All data 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 label and data sample is listed in Table 3.

Table 3. AUTO corpus Statistical tables

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

4.2 Evaluation Metrics

Following the previous work, we use Hamming-loss [19] and Micro-F1 [20] which are the most commonly used indicators in multi-label classification tasks.

Hamming-Loss. You can evaluate the difference between the predicted result sequence and the actual label sequence for the data in the test set. The higher the similarity between the two sequences, the lower the value of Hamming-loss, which means the better the result.

$$\text{hamming-loss}(h) = \frac{1}{p} \sum_{i=1}^p \frac{1}{Q} |h(x_i) \Delta Y_i| \quad (10)$$

Where Δ represents the symmetric difference between two sets, which is used to measure the degree of difference between the two sets.

Micro-F1. This is a micro-average, based on the basic quantities in the binary classification problem including true negative number (TN), false negative number (FN), true positive number (TP), false positive number (FP) evaluation indicators. Firstly, we calculate the average of the basic quantities of all labels, and then use the average to calculate the performance indicators of the classification.

4.3 Experimental Details

In this paper, 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).

This experiment uses the pre-trained vehicle domain word vector model as word representation, for words that are not in the vocabulary, replace them with ‘unks’. In order to avoid the influence of the vehicle brand on the prediction result, this paper makes synonymous substitution of the description of the vehicle brand and the vehicle system, and also makes corresponding substitution of the figures in the complaint text. 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 [13], 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, 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. We use the beam search algorithm [21] to find the highest ranked prediction path at the inference time. This prediction paths ending with the ‘eos’ are

added to the set of candidate paths. The length of the beam search is set to 5. In the training process, Adam optimizer is used to minimize the cross-entropy loss function.

4.4 Baselines

In order to compare 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 [11]: 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) [13]: 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 [22], and use Support Vector Machine (SVM) as the basic classifier in these algorithms.

4.5 Experimental Results and Analysis

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 4, Figs. 2 and 3, 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.

Table 4. Label prediction results comparison

Corpus	AUTO		DPAC	
	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

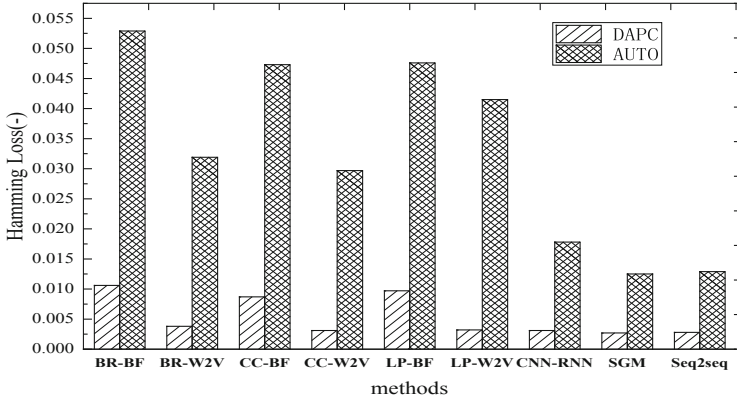


Fig. 2. Comparison of hamming loss

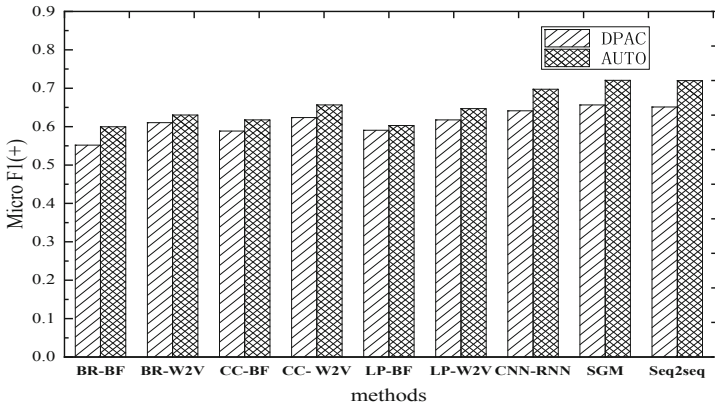


Fig. 3. Comparison of Micro-F1

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 5, Figs. 4 and 5.

Table 5. Label prediction results comparison

Corpus	AUTO		DPAC	
	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

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

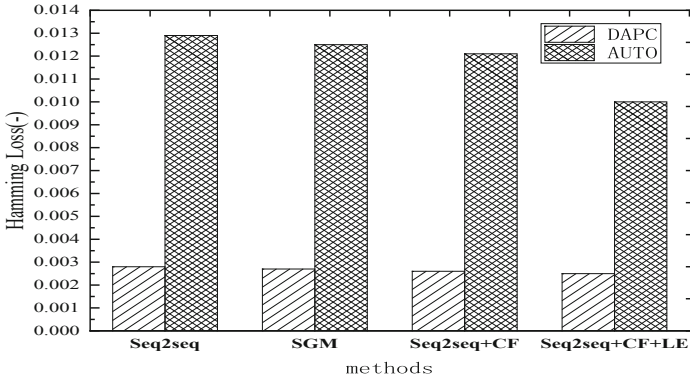


Fig. 4. Comparison of Hamming-loss

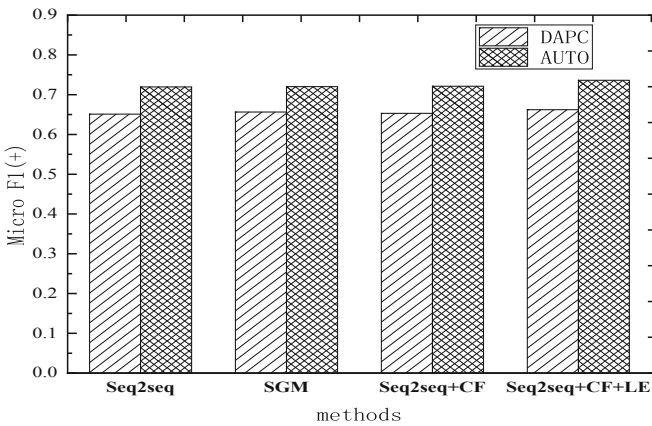


Fig. 5. Comparison of Micro-F1

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 the 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 6 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 according to the 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 6. Multi-label classification instances

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, don't understand the car can hear, and the car jitter, go to the store to check, say what is normal, jitter is normal.	Abnormal engine noise Body Vibration	Abnormal engine noise	Abnormal engine 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

5 Conclusion and Future Work

The multi-classification task of vehicle complaint data is of great significance in the process of vehicle defect recall. In this paper, we propose a multi-label classification method based on seq2seq model named VMIF-M to generate the defect label of vehicle complaint data. Firstly, the synonymous extension of defect description is made based on the existing defect classification system and the corpus related to vehicle complaint is collected to train a word vector model of vehicle domain. Then the word vector and defect label feature splicing are used as the input of the encoder, and then the encoder and decoder are connected through attention mechanism to focus on the words closely related to the defect label. Finally, the label distribution vector is added to the decoder, and the final classification prediction result is obtained through the softmax layer. This method avoids a lot of manual data processing. Experimental results show that the proposed methods outperform the baselines. The Macro-F1 reached 73% and 66% on the AUTO corpus and DPAC corpus, respectively. Through the analysis of the experimental data, we notice that the quality and size of the defect label library have a great influence on the prediction results. In the future work, the standardization of the vehicle defect in the process of vehicle recall will be used to improve the identification results of the complaint data.

Acknowledgments. This work is supported by National Natural Science Foundation of China under Grants No. 61671070, National Science Key Lab Fund project 6142006190301, National Language Committee of China under Grants ZDI135-53, and Project of Three Dimension Energy Consumption Saving Strategies in Cloud Storage System in Promoting the Developing University Intension–Disciplinary Cluster No. 5211910940.

References

1. Defective Product Administrative Center Homepage. <http://www.dpac.gov.cn>. Accessed 24 Jan 2019
2. Boutell, M.R., Luo, J., Shen, X.: Learning multi-label scene classification. *Pattern Recogn.* **37**(9), 1757–1771 (2004)
3. Zhang, M.L., Wu, L.: Lift: multi-label learning with label-specific features. In: *International Joint Conference on Artificial Intelligence*, pp. 1609–1614. AAAI Press (2017)
4. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. *Int. J. Data Warehous. Min.* **3**(3), 1–13 (2006)
5. Read, J., Pfahringer, B., Holmes, G.: Classifier chains for multi-label classification. *Mach. Learn.* **85**(3), 333 (2011)
6. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: De Raedt, L., Siebes, A. (eds.) *PKDD 2001. LNCS (LNAI)*, vol. 2168, pp. 42–53. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44794-6_4
7. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification (2002)
8. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recogn.* **40**(7), 2038–2048 (2007)
9. Li, L., Wang, H., Sun, X., et al.: Multi-label text categorization with joint learning predictions-as-features method. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 835–839 (2015)
10. Zhang, M.L., Zhou, Z.H.: Multi-label neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. Data Eng.* **18**(10), 1338–1351 (2006)
11. Chen, G., Ye, D., Xing, Z., et al.: Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In: *International Joint Conference on Neural Networks*, pp. 2377–2383. IEEE (2017)
12. Baker, S., Korhonen, A.: Initializing neural networks for hierarchical multi-label text classification. In: *BioNLP*, pp. 307–315 (2017)
13. Yang, P., Sun, X., Li, W., et al.: SGM: sequence generation model for multi-label classification (2018)
14. Chinese synonyms Toolkit. <https://github.com/huyingxi/Synonyms>. Accessed 24 Jan 2019
15. Chinese Word Segmentation Tool. <https://pypi.org/project/jieba/>. Accessed 24 Jan 2019
16. Zhang, Y., Li, B., Lv, X.: Research on domain term dictionary construction based on Chinese Wikipedia. *Image Processing, Computing and Big Data* (2018)
17. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vectorspace. *arXiv preprint arXiv:1301.3781* (2013)
18. Graves, A.: 2005 Special Issue: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Elsevier Science Ltd. (2005)
19. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* **37**(3), 297–336 (1999)

20. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
21. Wiseman, S., Rush, A.M.: Sequence-to-sequence learning as beam search optimization. CoRR, abs/1606.02960 (2016)
22. Szymański, P.: A scikit-based python environment for performing multi-label classification. arXiv preprint [arXiv:1702.01460](https://arxiv.org/abs/1702.01460) (2017)