ABR-HIC: Attention Based Bidirectional RNN for Hierarchical Industry Classification

Rongzhe Wei
Qian Xuesen College, School of Mathematics and Statisitcs
Xi'an Jiaotong University
Xi'an, China
jessonwrz@163.com

Bo Dong
National Engineering Lab of Big Data Analytics,
College of Distance Education
Xi'an Jiaotong University
Xi'an, China
dong.bo@mail.xjtu.edu.cn

Kuanzheng Yang
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
1329750315@qq.com

Huan He
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
hehuan@mail.xjtu.edu.cn

Qinghua Zheng
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
qhzheng@mail.xjtu.edu.cn

Jianfei Ruan
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
xjtu_jfruan@163.com

***Abstract*—Accurate industry classification of national economic activities as an important component in the construction of economic structure and as the basis of the formulation of economic policies and management of national economic activities has been gaining increasing attention. However, owing to the rapid growth in the number of industries, it is become increasingly difficult for tax bureaus to classify the registered taxpayers' industries. Conventional industrial classification methods only focus on the text features, which can not be analyzed and judged comprehensively according to the registration information, and can only carry on single-label classification since they neglect the primary and secondary relationships between the main and subsidiary industries, which can not meet application requirements. To better address these challenges, this paper proposes a model known as attention based bidirectional RNN for hierarchical industry classification (ABR-HIC), which is the first approach, to the best of our knowledge, to simultaneously address comprehensive registration information utilization and multi-label classification for the main and subsidiary industries. Our architecture establishes a bidirectional RNN using a word-attention mechanism, which is able to capture and fully utilize the text and non-text registration information for feature representation. By separating the taxpayer's primary and secondary multi-label classification problem corresponding to the main and subsidiary industries, respectively, into two subtasks and through multi-task learning, our model can provide comprehensive primary and secondary multi-industrial labels. Experiments were conducted on real tax data-sets of the Shaanxi Province, China and the results demonstrate the outstanding performance of our architecture in terms of both the classification effect and training time compared with those of state-of-the-art approaches.**

***Index Terms*—Hierarchical Industry Classification, Multi-task Learning, Multi-label Classification, Bidirectional RNN**

I. Introduction

Industry classification based on economic taxonomy aims to categorize companies into industrial groupings according to all-rounded registration information [1]. With the continuous development of the social economy and the rapid emerging of industries [19], the formulation of economic policies and the management of national economy increasingly depend on the accurate statistics of economic activities, which must be more finely classified according to the standardized industrial classification [20] [2]. Under these circumstances, the accurate industrial classification for national economic activities, which plays a key role in the construction of economic structure, has been arousing extensive interest. As industrial classification is of prominent significance for national macro-management in a wide range of fields, such as statistics, planning, finance, taxing, and employment, an accurate and comprehensive criterion for industry classification is indispensable. Since the launch of the *International Standard Industrial Classification of All Economic Activities* by the United Nations, unprecedented changes have taken place in the economic structure of numerous countries. The emergence of new technologies and divisions of labor among different organizations has led to the formation of new types of activities and industries, which pose a serious challenge on the statistics providers as well as the users. Facilitated by the comprehensive and detailed registration information of taxpayers, the effective utilization of text and non-text features remains a critical issue. In addition, with the surge in the number of industries, the business scope of many taxpayers is not limited to one primary industry and they could own subsidiary industries. In this case, the accurate classification of both primary industry and subsidiary industries simultaneously according to taxpayers is of great significance.

Conventional economic industry classification methods are either based on the prior knowledge of the staff to manually classify industries or the use of deep learning technics, which extract the text feature of the registration information of taxpayers, to automatically classify industries. The manual classification method acquires detailed information of the business scope from the taxpayers when they register to classify

the primary and subsidiary industries. However, as the staff usually lacks field expertise, they can only classify industries according to minor prior knowledge and their understanding of the registered information, this results in a lot of uncertainties and subjective judgment in the whole process. As a result, the overall accuracy for industrial classification is limited. Moreover, the labor cost of performing manual classification is very high. In contrast, the automatic classification methods only focus on the text features and neglect abundant non-text features, which can not be analyzed and judged comprehensively according to the registration information of the taxpayers and these methods can only predict one most-likely economic industry, which are not able to make accurate predictions when the taxpayers' business scope have more than one economic industry. Thus, the comprehensive and effective utilization of text and non-text features and the all-rounded multi-label predictions of economic industries with high accuracy remains a crucial issue.

In this paper, we propose an architecture termed attention based bidirectional RNN for hierarchical industry classification (ABR-HIC), which exploits the comprehensive utilization of text and non-text features for better classification accuracy and uses a multi-task learning method to effectively solve the problem of primary and secondary industrial label prediction. Specifically, our architecture establishes a fusion bidirectional RNN with word attention mechanism, which can fully utilize the text and non-text registration information for feature representation and capturing the underlying relations among features from different perspectives. For multi-industry prediction, we separated the primary and secondary multi-label classification problem of taxpayers corresponding to the main and subsidiary industries into two sub-tasks, and through multi-task learning, we can obtain comprehensive primary and secondary multi-industrial labels of taxpayers.

The proposed model is trained on real-tax dataset of the Shaanxi Province in China. Comprehensive experimental results demonstrated that our proposed architecture outperformed the state-of-the-art methods in classifying primary and secondary multi-label industries in terms of both the overall industry classification accuracy and comprehensiveness of multi-industry predictions. Furthermore, our proposed architecture can be applied not only to multi-industry classification, but also other multi-label classification problems.

The remainder of this paper is structured as follows. Section 2 summarizes the related work regarding the text-classification algorithm as well as multi-task learning methods. In section 3, we propose an architecture for multi-industry classification. Thereafter, we describe our tax dataset, experiments, and results in section 4. Section 5 presents the conclusion of our work.

II. Related Work

A. Text Classification Algorithm

The research on economic industrial classification actually focuses on the classification of the text of taxpayer's business scope. Existing studies have mostly adopted machine learning or deep learning methods for economic industrial classification according to the taxpayer's business scope. In addition, RNN [16] is adopted to classify the industry with the highest probability value as the classification result, however this method can only predict one industry and disregards the fact that in a real scenario, many taxpayers are likely to have more than one economic industry. Joulin et al. proposed FastText [3], which performs softmax classification according to the average word vector and uses n-gram features [17] to capture the local sequential information, which boosts the speed for model training and label predictions. Kim et al. proposed TextCNN [4], which uses the convolutional neural network (CNN) for local sequential information extractions. Although TextCNN achieves satisfying results in multi-label text classification, it can not model for longer or full-text sequential information. Liu et al. propose TextRNN [5] based on RNN, which can realize full-text sequential information modeling. TextBiRNN as an improved version of TextRNN, it changes the RNN layer into a bidirectional RNN layer, which can not only consider the forward coding information, but also the backward coding information. Raffel et al. propose TextAttBiRNN [6], which adds attention mechanism [7] to TextBiRNN and can capture the most relevant information for decision making. Yang et al. proposed hierarchical attention network (HAN) [8], which utilizes the hierarchical relationship among documents, sentences and words for layer-by-layer feature extractions [18]. However, in the economic industry classification, the taxpayer's business scope text is considerably different from the news text and other data sets that are often analyzed in the text classification. The business scope text usually does not have complete subject, predicate, object, table language and other syntactic structures, but only a brief description of the business content. Furthermore, the proposed methods on economic industrial classification only consider the text features of the taxpayer's registration information and disregard the non-text features. To the best of our knowledge, our proposed architecture comprehensively utilizes the text as well as non-text features of the registration information and effectively merges them into our network.

B. Multi-task Learning

In general, multiple loss functions are optimized simultaneously in multi-task learning. Even if this type of learning only optimizes only one loss function, it can also use auxiliary tasks to improve the effect of the original task model [22]. Therefore, Caruana proposes that multi-task learning can be improved by using specific domain information in relevant task training [9]. Long et al. proposed deep relation network [10] to improve the multi-task learning model based on computer vision by adding matrix priors in a fully connected layer. This network can learn the relationship among different tasks. Hashimoti et al. proposed joint many-task model [11] that separates tasks in three levels: word, syntactic, and semantic levels. Higher-level learning tasks depend on the output of lower-level learning tasks. To better capture the relative weights of different learning tasks, Ruder et al. proposed the

sluice network [12] to learn which layers and subspaces can share representation information to achieve a satisfying performance. According to the aforementioned methods, multi-task learning can not only train multi-tasks simultaneously, but also utilizes the shared representation information to improve a model's overall generalization performance. In this case, the proposed ABR-HIC model based on the multi-task learning method simultaneously trains the subtasks for single label classification, and thus efficiently and accurately solves the classification problem of primary and secondary multi-label of taxpayers' economic industries.

III. Proposed Architecture

A. Feature Extractions

The proposed ABR-HIC architecture comprehensively combines the text and non-text features, which are of significant importance in taxpayer's economic industrial classification. The details are presented as follows:

1) Text Feature Extractions: Our architecture adopts the most commonly used text features for industry classification, which are the taxpayer name and business scope. The taxpayer name is the enterprise name of the registered taxpayer, and its composition is shown in Table I. To reduce the influence of irrelevant content on the classification of taxpayers' economic industries, we deleted the administrative division component from the taxpayers name during text segmentation by constructing the corresponding dictionary of administrative division words. Thus, the manufacturer name, operational characteristics and organization form are selected as the sub-characteristics of the taxpayer-name feature. The business scope as a requisite in registration information reflects the primary and secondary economic activities that taxpayers engage in. However, as the business-scope text is usually brief and contains jargon of numerous industries, our architecture constructs a professional economic industry dictionary to improve the accuracy of word segmentation. To obtain the vectorization representation of text features, we adopt the Word2Vec [13] technic to convert the taxpayer-name and business-scope features into vectors.

2) Non-text Feature Extractions: Derived from specific scenario requirements, we comprehensively selected 14 non-text distinguishing features from the following four information categories: legal entity, capital operation, staff size, and type symbol. Nine of the fourteen features are quantitative characteristics, while the remaining are qualitative characteristics. The specific non-text features chosen are shown in Table II. However, the collected features have different degrees of data loss, quantitative indicators are not dimensionless, and qualitative indicators were not effectively coded. In this case, these features must be processed before inputting them into the classification model. For the processing of nine quantitative indicators, we first completed the missing data by using the zero padding method to ensure that all taxpayers' corresponding features are not null. In the process of nondimensionalization, we adopted the Z-score for the standardization of quantitative features:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where z represents the feature value after standardization, x represents the feature value before standardization, μ indicates the mean value, and σ denotes the standard deviation.

For the processing of five qualitative indicators selected, we filled in the missing data and used one-hot encoding [21] to vectorize the qualitative features in taxpayer's business pattern.

B. Task Decomposition of the Classification of Primary and Secondary Labels

In the task of classifying primary and secondary industrial labels, the primary and secondary industry labels of taxpayers are denoted as $\mathcal{Y}_p$ and $\mathcal{Y}_s$, respectively. Further, $\mathcal{Y}_o$ denotes the union of $\mathcal{Y}_p$ and $\mathcal{Y}_s$. As some taxpayers only have primary industries, their secondary industry labels will be vacant. According to our real-tax dataset, 82% samples of the taxpayers' secondary industry label are null, and thus the secondary labels are not suitable to be trained and predicted through deep learning models. Therefore, we decomposed the overall industry-classification task into subtasks of the classifications of taxpayers' total labels and primary labels, denoted as $Subtask_0$ and $Subtask_p$ respectively. $Subtask_0$ takes text features $\mathcal{X}^T$ and non-text features $\mathcal{X}^{NT}$ of taxpayers' samples as the input of the multi-label classification task, and it is defined as:

$$Subtask_o: \ Sigmoid\left(f_o\left(\mathcal{X}^T, \mathcal{X}^{NT}\right)\right) \Rightarrow \mathcal{Y}_o \tag{2}$$

where f_o is the mapping function of $Subtask_o$ in the multi-task learning model.

Similarly, $Subtask_p$ classification uses the same text and non-text features of $\mathcal{X}^T$ and $\mathcal{X}^{NT}$, respectively, of taxpayers' samples as the input, however, the subtask changes from multi-label classification to single-label classification, which the predicted result represents for the primary industry label. This $Subtask_p$ is defined as:

$$Subtask_p: \ Softmax\left(f_p\left(\mathcal{X}^T, \mathcal{X}^{NT}\right)\right) \Rightarrow \mathcal{Y}_p \tag{3}$$

where f_p is the mapping function of $Subtask_p$.

The two aforementioned subtasks observe and learn the primary and secondary multi-label classification tasks of the taxpayer industry from the comprehensive and key perspectives.

C. ABR-HIC Architecture

1) Architecture Framework: Figure 1 illustrates the proposed ABR-HIC architecture to perform multi-task learning on $Subtask_0$ and $Subtask_p$. The whole architecture is composed of four parts: the input layer, shared layer, task-specific layer, and output layer. Among them, the input layer takes text and non-text features as the input vectors and is shared by both $Subtask_0$ and $Subtask_p$. The shared layer is also shared by both two subtasks. For the task-specific layer, $Subtask_0$ and $Subtask_p$ own different network structures, and the output

TABLE I: Illustration of Taxpayer Name Composition

Administrative Division	Manufacturer Name	Operational Characteristics	Organization Form	Economic Industry
Xi'an City	Bie Ju Yi Ge	Network Technology	Company Limited	Software Development
Yuyang District, Yulin City	Rong Sheng	Vehicle Maintenance	Center	Vehicle Repair
Dingbian County	jiang Ping	Ceramics	Direct-Sale Store	Stone Materials Retail

TABLE II: Taxpayer Non-text Features

Feature Type	Feature Number	Feature Symbol	Feature Definition	Data Type
Legal Entity Information	1	FDDBR_AGE	Legal Representative Age	Quantitative
	2	FDDBR_SEX	Legal Representative Sex	Qualitative
Capital Operation Information	3	ZCZB	Registered Capital	Quantitative
	4	ZRR_TZBL	Proportion of Natural Person Investment	Quantitative
	5	WZ_TZBL	Proportion of Foreign Investment	Quantitative
	6	GY_TZBL	Proportion of State-owned Investment	Quantitative
Staff Size Information	7	CYRS	Employee Number	Quantitative
	8	WJRS	Foreigner Number	Quantitative
	9	HHRS	Partnership Number	Quantitative
	10	GDRS	Fixed Number	Quantitative
Type Symbol Information	11	DJZCLX	Registration Type	Qualitative
	12	JYFS	Business Pattern	Qualitative
	13	ZJGBZ	Head Office Logo	Qualitative
	14	GGHBZ	National Land Tax Co-administrators	Qualitative

layer predicts the total labels or the primary label of taxpayers' industries.

Specifically, $Subtask_0$ and $Subtask_p$ share the word encoder part, and this results in the same text-feature coding results on these two relevant subtasks, because both subtasks use text feature $\mathcal{X}^T$, and a large amount of time is required during the training process of word embedding and bidirectional RNN. Moreover, as there is relatively strong correlation between the two subtasks, their text-feature coding requirements are consistent. Therefore, the modeling of text features of two subtasks is accomplished by sharing the network structure, which can save time in the training process, and the sharing mechanism of hidden-layer parameters is used to improve the generalization performance of the text-feature modeling. The task-specific layers include exclusive network structures of $Subtask_o$ and $Subtask_p$. Both the exclusive networks contain word attention layers, feature combination layers, and fully connected layers. The word attention layers will extract word attention corresponding to the training target of each subtask. In feature combination layers, both the word attention extracted from word attention layers and non-text features $\mathcal{X}^{NT}$ are concatenated. Finally, the fully connected layers integrate the classified local information contained in the concatenated distributed features. For the output layer of $Subtask_0$, we set 0.5 as the threshold value for all predicted taxpayers's industry labels. As the softmax layer of $Subtask_p$ accomplishes the predictions on all labels, we mapped the results into $[0, 1]$ interval, and selected the label with the maximum probability as the primary industry label.

2) Joint Loss: According to the research by Nam et al. [15], the use of cross entropy loss function in large-scale multi-label classification can achieve a very satisfying classification result. In this case, we selected the cross entropy loss function for both $Subtask_0$ and $Subtask_p$. The cross entropy loss function for a single sigmoid neuron is defined as:

$$\mathcal{L}^{si} = -\frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{Y}_i \log \hat{\mathcal{Y}}_i + (1-\mathcal{Y}_i)\log\left(1-\hat{\mathcal{Y}}_i\right)\right) \quad (4)$$

where $\mathcal{Y}$ represents the original label, $\hat{\mathcal{Y}}$ denotes the probability of a single input sample, and N is the total number of samples. In $Subtask_0$, the sigmoid layer contains M neurons,

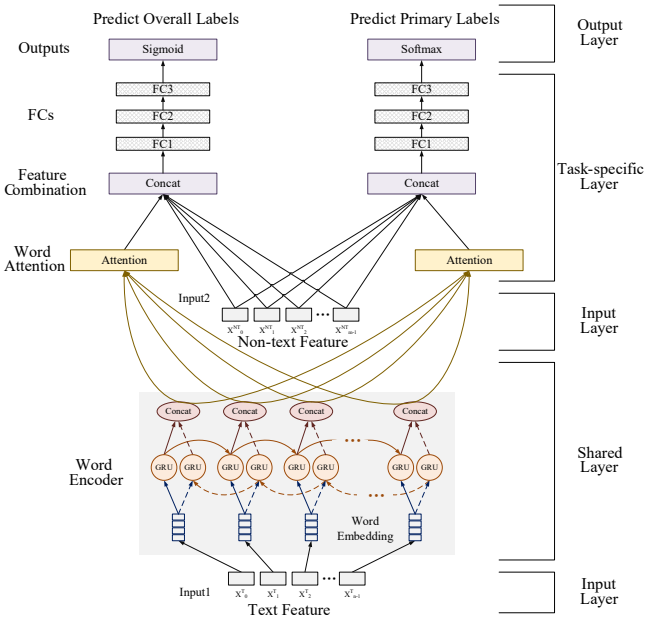

Fig. 1: ABR-HIC architecture.
Illustration of the proposed attention based bidirectional RNN for hierarchical industry classification (ABR-HIC) architecture.

and each neuron corresponds to a single industry label. Thus, the total loss of the whole subtask is defined as:

$$\begin{aligned}&\mathcal{L}\left(Sigmoid\left(f_o\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_o\right)\right)=\frac{1}{M}\sum_{j=1}^{M}\mathcal{L}_j^{si}\\&=-\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{Y}_{ij}\log\hat{\mathcal{Y}}_{ij}+(1-\mathcal{Y}_{ij})\log\left(1-\hat{\mathcal{Y}}_{ij}\right)\right)\end{aligned}\tag{5}$$

For $Subtask_p$, the total loss is defined as:

$$\begin{aligned}&\mathcal{L}\left(Softmax\left(f_p\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_p\right)\right)\\&=-\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\mathcal{Y}_{ij}\log\hat{\mathcal{Y}}_{ij}\end{aligned}\tag{6}$$

As both loss functions are calculated distinctively, their values are in different orders of magnitude. Therefore, we propose a dynamic loss relative weight to balance the different loss functions for different tasks according to the orders of magnitude. By the end of each epoch, our model calculates the relative weights dynamically according to loss values from different tasks. The loss relative weights for $Subtask_0$ and $Subtask_p$ are denoted as W_o and W_p respectively, and formulated as:

$$W_o=\frac{\mathcal{L}\left(Soft_p\right)}{\mathcal{L}\left(Sigmoid_o\right)+\mathcal{L}\left(SoftmaxSoft_p\right)}\tag{7a}$$

$$W_p=\frac{\mathcal{L}\left(Sigmoid_o\right)}{\mathcal{L}\left(Sigmoid_o\right)+\mathcal{L}\left(SoftmaxSoft_p\right)}\tag{7b}$$

$$Sigmoid_o=Sigmoid\left(f_o\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_o\right)\tag{7c}$$

$$Soft_p=Softmax\left(f_p\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_p\right)\tag{7d}$$

As shown in equations $(7a)-(7d)$ if the loss value is relatively large, its loss weight will be relatively small, and thus, the equations guarantee a balance between different loss functions.

In multi-task learning, as different task weights can influence the training results, the joint loss for multi-task learning can be defined as:

$$\begin{aligned}\mathcal{L}_{joint}=&\lambda W_o\mathcal{L}\left(Sigmoid\left(f_o\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_o\right)\right)\\&+(1-\lambda)\left(1-W_o\right)\mathcal{L}\left(Softmax\left(f_p\left(\mathcal{X}^T,\mathcal{X}^{NT}\right),\mathcal{Y}_p\right)\right)\end{aligned}\tag{8}$$

where λ is the task weight for $Subtask_0$ and is evaluated from 0 to 1, and $1-\lambda$ is the task weight for $Subtask_p$. Thus, the main goal for multi-task learning is to minimize $mathcalL_{joint}$ to achieve the best fitting results.

3) Primary and Secondary Label Task Mergence: As $Subtask_0$ and $Subtask_p$ are decomposed from the primary and secondary label tasks, respectively, we propose a label-aggregation reordering algorithm to integrate the results of the subtasks and obtain the overall industry-classification results. The pseudo-code of this label-aggregation reordering algorithm is shown in Algorithm **1**.

Algorithm 1 Label Aggregation Reordering Algorithm

Require: N - Number of samples; M - Number of labels; $\mathcal{Y}$ - Original labels; $\mathcal{P}$ - Prediction probability; Th - Threshold of probability;
Ensure: $\mathcal{Y}_p$ - Predicted Labels;
0: $\mathcal{Y}_p=[\,]$; // init the predicted labels
0: **for** $i=0:N$ **do**
0: $L_o=[\,]$; // init the predicted labels for the sample[i]
0: **while** $\mathcal{P}_o[i][\arg\max(\mathcal{P}_o[i])]>0.5$ **do**
0: $L_o.append\left(\arg\max\left(\mathcal{P}_o[i]\right)\right)$; // append overall labels in order
0: $\mathcal{P}_o[i][\arg\max(\mathcal{P}_o[i])]=0$;
0: **end while**
0: $L_p=\arg\max\left(\mathcal{P}_p[i]\right)$; // the primary label
0: $L_m=[\,]$; // init merged labels
0: $L_m.append\left(L_p\right)$;
0: **for** $k\leftarrow L_o$ **do**
0: **if** ! $L_m.contains\left(k\right)$ **then** $L_m.append\left(k\right)$;
0: **end if**
0: **end for**
0: $\mathcal{Y}_p.append\left(L_m\right)$;
0: **end for**
0: **return** $\mathcal{Y}_p$; =0

IV. Experimental Results

To prove the effectiveness of the proposed ABR-HIC architecture, we first introduce our dataset and metrics. Then, we investigate the following research questions:

Question 1: Can our ABR-HIC architecture achieve high classification accuracy than state-of-the-art approaches? For this, we limited our model to single-label-classification tasks with only text-feature or multi-feature as input, and we compared the selected metrics using both the proposed and state-of-the-art methods to prove its effectiveness.

Question 2: How effective is the multi-task learning method in classifying taxpayer industries? To illustrate this, we evaluated and compared the classification results of the multi-task learning method with those of single-task learning in terms of both training time and overall classification accuracy.

A. Dataset and Metrics

1) Dataset: The dataset of the experimental evaluation in this study adopts the real taxpayer-registration information obtained from the tax bureau of the Shaanxi province in China. To ensure that the samples selected for the experiments adopt the same industrial classification standard, we selected taxpayers whose registration time falls within the period of November 1, 2011 to September 30, 2017, and classified them into economic industries according to the revised edition of the standard of *Industrial Classification for National Economic Activities*, 2011. From the selected samples, we sorted 86 industries for training samples, ensuring that every industry category owns no less than 1000 samples. In the case of an industry category with more than 5000 samples, to maintain a relative balance among categories, we used random under-sampling. After processing, we split the samples into training, validation, and test sets in the ratio of $8:1:1$. The numbers for each set are shown in Table III.

TABLE III: Sample Partition Table

Total	Training Set	Validation Set	Test Set
271116	216838	27141	27137

A taxpayer can have a primary industry and up to three secondary industries, thus, we calculated the statistics of the quantity of labels on samples. These statistical results of different labels and sample cumulative percentage are shown in Figure 2.

2) Metrics: According to the summary of multi-label classification by Zhang et al. [14], we selected nine classification indicators, which can be divided into sample and category-based indicators. Sample-based indicators include Precision

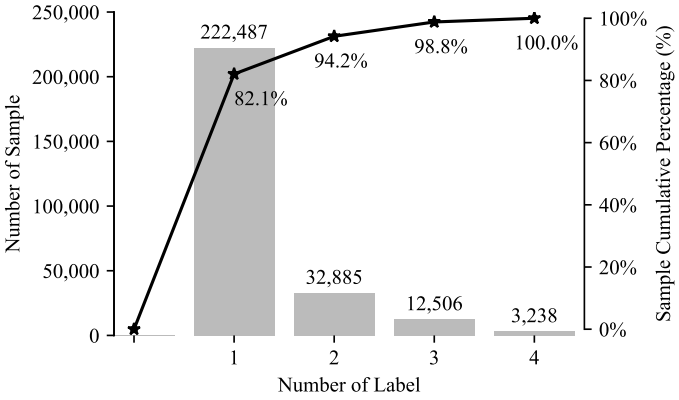

Fig. 2: Quantity of Labels on Sample and Sample Cumulative Percentage

$Precision_{exam}$, Recall $Recall_{exam}$, and F-Measure F^{β}_{exam}, and are defined as follows:

$$Precision_{exam} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap H(x_i)|}{|H(x_i)|} \tag{9a}$$

$$Recall_{exam} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap H(x_i)|}{|Y_i|} \tag{9b}$$

$$F^{\beta}_{exam} = \frac{\left(1+\beta^2\right)\cdot Precision_{exam}\cdot Recall_{exam}}{\beta^2\cdot Precision_{exam} + Recall_{exam}} \tag{9c}$$

where Y_i represents the actual label set to which the i^{th} sample belongs, and H_i denotes the predicted label set of the i^{th} sample.

By calculating F-Measure for result evaluation, we set $\beta = 1$. In this case, $F1_{exam}$ value is used as the harmonic average of Precision and Recall to measure the comprehensive effect of model classification.

Category-based indicators include Precision $Precision$, Recall $Recall$, Macro-average B_{macro} and Micro-average B_{micro} of F-Measure value. For the j^{th} category, we divided the samples into four parts, i.e., TP_j, FP_j, TN_j, and FN_j corresponding to true positive, false positive, true negative, and false negative samples, respectively.

$$TP_j = |\{x_i | y_j \in Y_i \wedge y_j \in H(x_i), 1 \le i \le N\}| \tag{10a}$$

$$FP_j = |\{x_i | y_j \notin Y_i \wedge y_j \in H(x_i), 1 \le i \le N\}| \tag{10b}$$

$$TN_j = |\{x_i | y_j \notin Y_i \wedge y_j \notin H(x_i), 1 \le i \le N\}| \tag{10c}$$

$$FN_j = |\{x_i | y_j \in Y_i \wedge y_j \notin H(x_i), 1 \le i \le N\}| \tag{10d}$$

For the j^{th} category, $B(TP_j, FP_j, TN_j, FN_j)$ denotes the

Precision, Recall, and F-Measure, which are defined as:

$$Precision\,(TP_j, FP_j, TN_j, FN_j) = \frac{TP_j}{TP_j + FP_j} \quad (11a)$$

$$Recall\,(TP_j, FP_j, TN_j, FN_j) = \frac{TP_j}{TP_j + FN_j} \quad (11b)$$

$$F^{\beta}\,(TP_j, FP_j, TN_j, FN_j) \quad (11c)$$

$$= \frac{\left(1+\beta^2\right)\cdot TP_j}{(1+\beta^2)\cdot TP_j + \beta^2 \cdot FN_j + FP_j} \quad (11d)$$

Accordingly, we can calculate the Macro-average B_{macro} and Micro-average B_{micro} as

$$B_{macro} = \frac{1}{p}\sum_{j=1}^{p} B\,(TP_j, FP_j, TN_j, FN_j) \quad (12)$$

$$B_{micro} = B\left(\sum_{j=1}^{p} TP_j, \sum_{j=1}^{p} FP_j, \sum_{j=1}^{p} TN_j, \sum_{j=1}^{p} FN_j\right) \quad (13)$$

In the F-Measure evaluation according to the category indicators, we set $\beta = 1$. In this case, $F1_{macro}$ and $F1_{micro}$ are used as the harmonic averages of Precision and Recall under the macro and micro-average conditions, respectively, to measure the comprehensive effect of model classification.

The above-mentioned indicators are all positive, thus, to achieve better industry-classification results the indicators must be maximized.

B. Effectiveness of ABR-HIC

1) Performance Calculation using Different Approaches: The experiments conducted were divided into two groups. In the first experiment, we used only the text features of taxpayers as input and neglected non-text features. In the second experiment, we used both text and non-text features simultaneously for training. In these experiments, we changed our architecture from multi-task learning to single-task learning for better comparison. To prove the effectiveness of our ABR-HIC architecture, we compared our method with state-of-the-art methods: (1) FastText, (2) TextCNN, (3) TextRNN, (4) TextBiRNN, (5) TextCRNN, (6) TextRCNN, and (7) TextHAN. In the case of the multi-input task in the second experiment, we changed the state-of-the-art methods into multi-input acceptance models: MI-FastText, MI-TextCNN, MI-TextRNN, MI-TextBiRNN, MI-TextCRNN, MI-TextRCNN, and MI-TextHAN models.

For these deep learning models, GRU layers were used to improve the speed of model training.

The classification results on the first experiment are shown in table IV. When utilizing text features for taxpayer economic industry classification, although the FastText model takes the least time for training, it performs poorly for all indicators. As TextCNN uses a CNN for local sequential information extractions, it performs better than FastText. The TextRNN model, which is able to realize full-text sequential information modeling, achieves a better score than the previous two architectures. As an improved version of TextRNN, TextBiRNN changes the RNN layer into a bidirectional RNN layer and performs slightly better than TextRNN. Although both TextCRNN and TextRCNN achieve satisfying classification results, their overall accuracy is slightly lower than that of our proposed architecture. In general, our ABR-HIC model is comprehensively superior to any other model. The attention mechanism adopted in our model can better capture the small amount of essential information from taxpayers' text features and ignore the most unimportant information; this boosts the classification accuracy. As the text features adopted in the classification of taxpayers' economic industry generally have a small number of words after word segmentation, which is not applicable to the multi-level attention model of words and sentences in TextHAN, the overall accuracy of the TextHAN model is lower than that of our proposed model.

Table V lists the classification results of the second experiment, and its shows that models training with both text and non-text features perform better than those training with only text features.

The result comparisons of $F1_{exam}$, $F1_{macro}$, and $F1_{micro}$ by using different methods are shown in Figure 3, Figure 4, and Figure 5, respectively. The figures not only demonstrate the effectiveness of utilizing non-text features for taxpayer industry classification but also the superiority of our method.

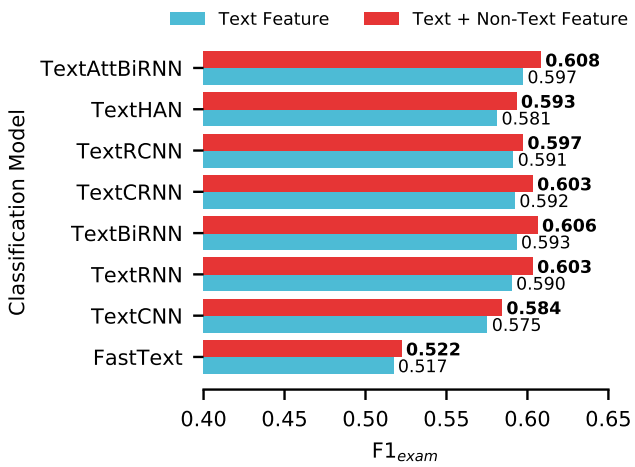

Fig. 3: Results on $F1_{exam}$

2) Performance on Multi-task Learning: To illustrate the effectiveness of the multi-task learning method adopted in our model, we evaluated the classification results of multi-task learning with those of single-task learning. As shown in the previous experiment, non-text features are effective for improving classification accuracy; therefore, we used both text and non-text features as input vectors. In single-task learning, the experiment that takes total labels $\mathcal{Y}_o$ of the taxpayers as the training target and selects the maximum probability label as the primary label is denoted as $STL_{overall}$, while the experiment that takes primary label $\mathcal{Y}_p$ as the training target is denoted as $STL_{primary}$. In the $STL_{overall}$ experiment, total-label-prediction probability P_o and primary-

TABLE IV: Experiment of Industry Classification Based on Text Features

Model	P_{exam}	R_{exam}	$F1_{exam}$	P_{macro}	R_{macro}	$F1_{macro}$	P_{micro}	R_{micro}	$F1_{micro}$
FastText	0.539	0.497	0.517	0.526	0.449	0.464	0.535	0.448	0.487
TextCNN	0.598	0.554	0.575	0.572	0.526	0.532	0.598	0.506	0.548
TextRNN	0.615	0.567	0.590	0.599	0.543	0.549	0.618	0.524	0.567
TextBiRNN	0.618	0.570	0.593	0.595	0.548	0.554	0.620	0.527	0.570
TextCRNN	0.615	0.571	0.592	0.594	0.546	0.551	0.618	0.527	0.569
TextRCNN	0.615	0.569	0.591	0.592	0.549	0.552	0.616	0.525	0.567
TextHAN	0.607	0.557	0.581	0.591	0.529	0.541	0.611	0.513	0.558
ABR-HIC	**0.621**	**0.575**	**0.597**	**0.601**	**0.551**	**0.557**	**0.624**	**0.533**	**0.575**

TABLE V: Experiment of Industry Classification Based on Text Features and Non-text Features

Model	P_{exam}	R_{exam}	$F1_{exam}$	P_{macro}	R_{macro}	$F1_{macro}$	P_{micro}	R_{micro}	$F1_{micro}$
MI-FastText	0.545	0.502	0.522	0.534	0.454	0.470	0.540	0.453	0.493
MI-TextCNN	0.607	0.564	0.584	0.588	0.532	0.542	0.606	0.518	0.559
MI-TextRNN	0.629	0.579	0.603	0.612	0.556	0.563	0.633	0.535	0.580
MI-TextBiRNN	0.630	**0.585**	0.606	0.612	**0.561**	0.565	0.633	0.543	0.584
MI-TextCRNN	0.624	0.583	0.603	0.609	0.557	0.566	0.626	0.542	0.581
MI-TextRCNN	0.621	0.575	0.597	0.602	0.553	0.559	0.624	0.531	0.574
MI-TextHAN	0.617	0.571	0.593	0.602	0.546	0.554	0.619	0.528	0.570
ABR-HIC	**0.632**	**0.585**	**0.608**	**0.622**	**0.561**	**0.572**	**0.635**	**0.544**	**0.586**

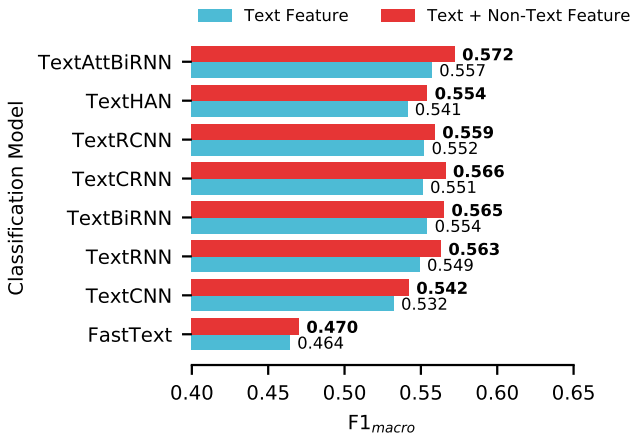

Fig. 4: Results on $F1_{macro}$

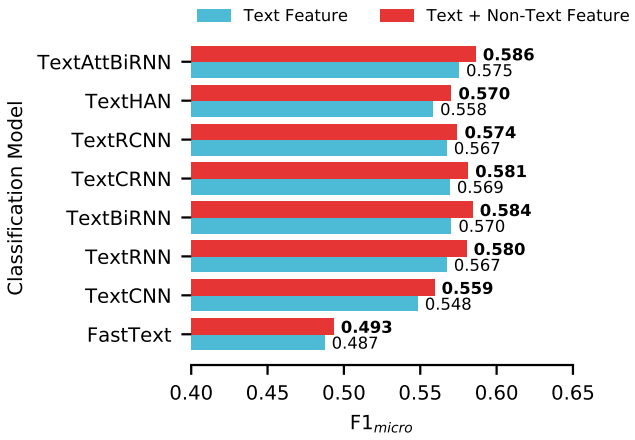

Fig. 5: Results on $F1_{micro}$

label-prediction probability P_p were combined according to the label-aggregation reordering algorithm to obtain the classification results by combining two single-learning tasks. Then, prediction and evaluation were carried out on total labels $\mathcal{Y}_o$ and primary label $\mathcal{Y}_p$, and this experiment was recorded as STL_{merge}. In multi-task learning, $Subtask_0$ and $Subtask_p$ were trained simultaneously. Let $MTL_{overall}$ denote the experiment on $Subtask_0$ that uses only one label with the largest probability prediction. The experiment on $Subtask_p$ is denoted as $MTL_{primary}$. The definition of MTL_{merge} is similar to that of STL_{merge} but under the multi-task learning condition. Single-task learning and multi-task learning were compared according to two aspects: training time and classification accuracy. Considering the training time, $STL_{overall}$ and $STL_{primary}$ are trained on $Task_o$ and $Task_p$ separately, where $Task_o$ and $Task_p$ represent the tasks for total- and single-label classifications, respectively. In multi-task learning, $Subtask_0$ and $Subtask_p$ are trained simultaneously. We calculated the time for every task, as shown in Figure 6. The figure reveals that the multi-task learning method utilizes 13% lesser training time than single-task learning.

For the classification accuracy with respect to total labels,

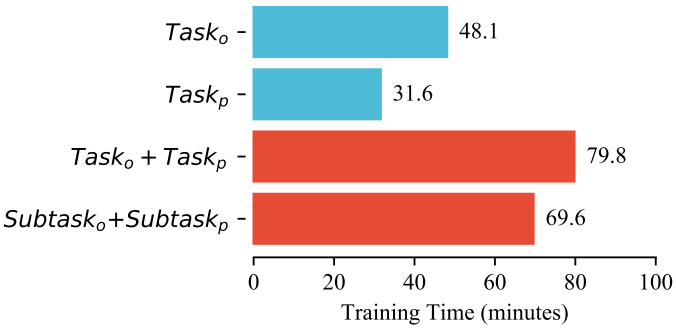

Fig. 6: Task Time Comparison

we calculated $F1_{exam}$, $F1_{macro}$, and $F1_{micro}$ for both multi-task and single-task learning (Figure 7). Similarly, we calculated $F1_{macro}$ and $F1_{micro}$ for both multi-task and single-task learning, and the results are shown in Figure 8.

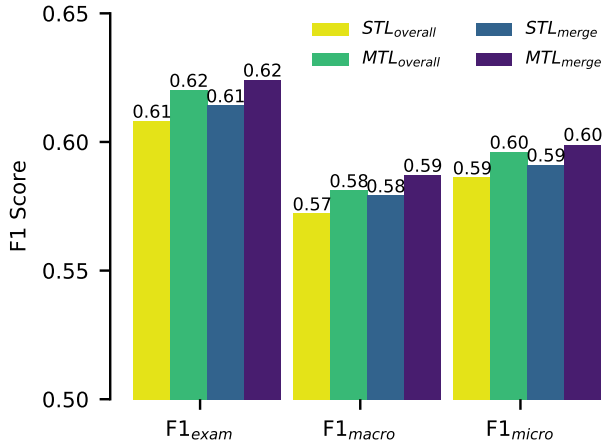

Fig. 7: Comparison Between Single-task Learning and Multi-task Learning on Total Label Classification

As illustrated in Figure 6, 7, and 8, multi-task learning method adopted in our model takes advantages in total label classification as well as single-label classification.

V. Conclusions

In this paper, we proposed an architecture named ABR-HIC for primary and secondary multi-label industry classification. Our architecture exploits the comprehensive utilization of text and non-text features for a better classification accuracy, establishes a fusion bidirectional RNN with word attention mechanism to capture the underlying relations among features from different perspective, and uses the multi-task learning method to effectively solve the primary and secondary industry label problem. The experimental results revealed that our architecture achieves outstanding performance in both the classification accuracy and training time than state-of-the-art models.

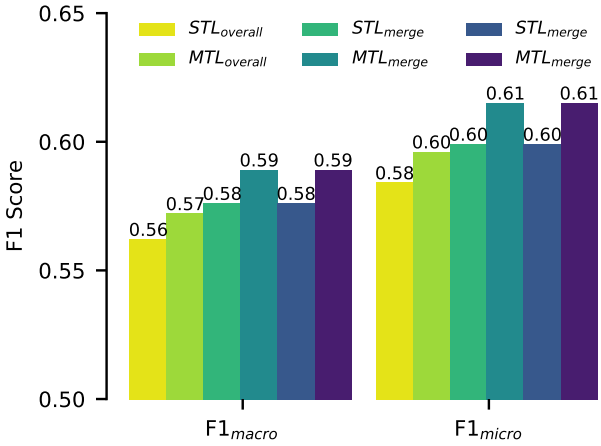

Fig. 8: Comparison Between Single-task Learning and Multi-task Learning on Primary Label Classification

Although we trained our architecture in the economic-industry-classification scenario, our model can be applied to other multi-label classification scenarios as well.

Acknowledgment

This research was partially supported by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with Grant No. 2018YFB1004500, the MOE Innovation Research Team No. IRT17R86, the National Science Foundation of China under Grant Nos. 61721002, 61532015, and Project of China Knowledge Centre for Engineering Science and Technology.

References

[1] P. J. Devine et al., An Introduction to Industrial Economics. Routledge, 2018.
[2] J. Pucher, Z. Peng, N. Mittal, Y. Zhu, and N. Korattyswaroopam, Urban Transport Trends and Policies in China and India: Impacts of Rapid Economic Growth, Transport Reviews, vol. 27, no. 4, pp. 379C410, Jul. 2007.
[3] [1607.01759] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, Bag of Tricks for Efficient Text Classification, arXiv:1607.01759 [cs], Jul. 2016.
[4] Y. Kim, Convolutional Neural Networks for Sentence Classification, arXiv:1408.5882 [cs], Aug. 2014.
[5] P. Liu, X. Qiu, and X. Huang, Recurrent Neural Network for Text Classification with Multi-Task Learning, arXiv:1605.05101 [cs], May 2016.
[6] C. Raffel and D. P. W. Ellis, Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems, arXiv:1512.08756 [cs], Dec. 2015.
[7] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473 [cs, stat], Sep. 2014.
[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, Hierarchical Attention Networks for Document Classification, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, 2016, pp. 1480C1489.
[9] R. Caruana, A Dozen Tricks with Multitask Learning, in Neural Networks: Tricks of the Trade, G. B. Orr and K.-R. Mller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 165C191.

[10] M. Long and J. Wang, Learning Multiple Tasks with Deep Relationship Networks, ArXiv, vol. abs/1506.02117, 2015.
[11] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher, A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks, arXiv:1611.01587 [cs], Nov. 2016.
[12] S. Ruder, J. Bingel, I. Augenstein, and A. S?gaard, Sluice networks: Learning what to share between loosely related tasks, ArXiv, vol. abs/1705.08142, 2017.
[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111C3119.
[14] M. Zhang and Z. Zhou, A Review on Multi-Label Learning Algorithms, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 8, pp. 1819C1837, Aug. 2014.
[15] [1]J. Nam, J. Kim, E. Loza Menca, I. Gurevych, and J. Frnkranz, Large-Scale Multi-label Text Classification Revisiting Neural Networks, in Machine Learning and Knowledge Discovery in Databases, 2014, pp. 437C452.
[16] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, Recurrent Neural Network Based Language Model, p. 4.
[17] M. Pagliardini, P. Gupta, and M. Jaggi, Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 528C540, 2018.
[18] Y. Jang, J. Ham, B. Lee, and K. Kim, Cross-Language Neural Dialog State Tracker for Large Ontologies Using Hierarchical Attention, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 11, pp. 2072C2082, Nov. 2018.
[19] S. A. Elnagdy, M. Qiu, and K. Gai, Cyber Incident Classifications Using Ontology-Based Knowledge Representation for Cybersecurity Insurance in Financial Industry, in 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), 2016, pp. 301C306.
[20] S. A. Elnagdy, M. Qiu, and K. Gai, Cyber Incident Classifications Using Ontology-Based Knowledge Representation for Cybersecurity Insurance in Financial Industry, in 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), 2016, pp. 301C306.
[21] D. Harris and S. Harris, Digital Design and Computer Architecture. Morgan Kaufmann, 2010.
[22] S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks, arXiv:1706.05098 [cs, stat], Jun. 2017.