electronics

MDPI

Article

DWSA: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining

Tan Yue, Yong Li and Zonghai Hu *

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; yuetan@bupt.edu.cn (T.Y.); yli@bupt.edu.cn (Y.L.)

* Correspondence: zhhu@bupt.edu.cn; Tel.: +86-010-62283467

Abstract: The structure of a document contains rich information such as logical relations in context, hierarchy, affiliation, dependence, and applicability. It will greatly affect the accuracy of document information processing, particularly of legal documents and business contracts. Therefore, intelligent document structural analysis is important to information extraction and data mining. However, unlike the well-studied field of text semantic analysis, current work in document structural analysis is still scarce. In this paper, we propose an intelligent document structural analysis framework through data pre-processing, feature engineering, and structural classification with a dynamic sample weighting algorithm. As a typical application, we collect more than 11,000 insurance document content samples and carry out the machine learning experiments to check the efficiency of our framework. Meanwhile, to address the sample imbalance problem in the hierarchy classification task, a dynamic sample weighting algorithm is incorporated into our Dynamic Weighting Structural Analysis (DWSA) framework, in which the weights of different category tags according to the structural levels are iterated dynamically in training. Our results show that the DWSA has significantly improved the comprehensive accuracy and the classification F1-score of each category. The comprehensive accuracy is as high as 94.68% (3.36% absolute improvement) and the Macro F1-score is 88.29% (5.1% absolute improvement).

Keywords: artificial intelligence; big data; data mining; intelligent document structural analysis; machine learning; sample imbalance

check for updates

Citation: Yue, T.; Li, Y.; Hu, Z. DWSA: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining. *Electronics* **2021**, *10*, 2443. https://doi.org/10.3390/electronics10192443

Academic Editors: Alberto Fernandez Hilario and Matteo Golfarelli

Received: 1 August 2021
Accepted: 28 September 2021
Published: 8 October 2021

1. Introduction

With the rapid growth of insurance policies, consumers and insurance industry practitioners need to effectively extract useful information from a vast number of insurance documents. For time and manpower consuming activities such as identifying the key terms and differences between similar policies, intelligent document processing methods are highly desirable. Various document processing techniques have been developed, allowing electronic documents to be processed more efficiently [1–4]. In recent years, machine learning methods have been widely used in data mining and information extraction and received great success. However, in the specific field of insurance industry, Chinese insurance documents pose extra challenges because most exist in unstructured forms and the text features are often mixed with some noise, leading to failure of feature recognition when processing documents. Therefore, further work is needed to automatically process unstructured insurance documents to extract key information and transform them into structured data without losing important features.

This paper proposes an intelligent information extraction framework for Chinese insurance documents. The proposed model can provide a convenient technology platform for insurance practitioners or related researchers. In addition, a dynamic sample weight algorithm is proposed to address the sample imbalance problem.

1.1. Related Studies and the Current Contribution

Machine learning methods are widely used in data mining and document analysis tasks [5–10]. Automatic classification systems have been applied to mail classification, electronic conference, information filtering, and so on [11–15]. The Xgboost algorithm [16] implements the GBDT algorithm and has been widely used in competitions and many other machine learning projects and achieved satisfying results [17,18]. The Lightgbm algorithm [19] supports efficient parallel training, which processes data quickly and also performs well. Gómez et al. [20] proposed a model that when given an initial UML model, a number of alternatives suitable for data structures can be generated semi-automatically. Liu. et al. [21] provided a comprehensive survey of the progress that has been made in the field of document image classification over the past two decades.

In terms of sample imbalance, many researchers use the methods of over-sampling and under-sampling to solve the problem of sample imbalance [22]. Lin et al. [23] proposed a new loss function: Focal loss. The loss function is modified based on the standard cross entropy loss. This function can make the model focus more on difficult samples in training by reducing the weight of easily classified samples.

To address the serious challenge of information explosion, automated tools are needed to help people quickly extract specific factual information from a large number of information sources. [24–28]. A lot of works have been done in the field of entity recognition and extraction, knowledge acquisition [29–32], and text analysis and reasoning, etc. [33–36]. Quan et al. [37] proposed a paper classification and information extraction system in the computer science field. The team used the Naive Bayes algorithm to automatically classify a large number of papers and extract relevant information. In terms of algorithm, a new Weighted Naive Bayes model is developed to better fit the data model. The data set only contains the paper abstract. The documents are not in PDF format and are relatively easy to process. Still, the accuracy of the information extraction system cannot reach the application standard.

Compared with English language processing techniques, Chinese language has a lot of characteristics different from English [38]. For example, there is no natural boundary between words in Chinese, so before the classification, the text should be divided into words. In addition, the proportion of syntactic analysis and semantic analysis is different in different languages [39,40]. Therefore, Chinese language processing needs additional data pre-processing work. At present, the main difficulty in Chinese document processing is that the accuracy or F1-score of intelligent recognition and classification is not high enough to achieve the target performance. As a result, the accuracy and efficiency of information extraction cannot achieve the application standard.

The whole task of Chinese insurance document analysis can be divided into the following parts: (1) Data pre-processing and data cleaning are conducted at first. Most insurance documents exist in PDF format, and all document contents need to be recognized and converted. (2) An algorithm model needs to be built for structural classification and information extraction. (3) Structured documents are the output, including the re-typesetting of the original text, tables, and other information in the document. Accordingly, we design the Dynamic Weighting Structural Analysis (DWSA) framework. Compared with related works, the proposed method is first applied in the Chinese insurance document field, and the dynamic sample weight algorithm can improve the performance of the structural classification and address the problem of sample imbalance effectively.

The main contributions of this paper are summarized as follows.

- This paper proposes an intelligent information extraction framework for insurance documents. We use the techniques of data pre-processing, feature engineering, and structural classification to process the documents. The proposed model can provide a convenient technology platform for insurance practitioners or related researchers.

- A dynamic sample weight algorithm is proposed to address the sample imbalance problem and improve the structural classification part of the DWSA framework. The comprehensive accuracy reaches 94.68% (3.36% absolute improvement). And the Macro F1-score achieves 88.29% (5.1% absolute improvement).

1.2. DWSA Framework

In this work, first an intelligent structural analysis (SA) framework is proposed to help people to survey a large amount of Chinese insurance documents. Secondly, a dynamic weighting (DW) classification algorithm based on Adaboost is proposed to address the sample imbalance problem and improve the classification part in the structural analysis framework. Through the structural analysis, the whole document contents are divided into several hierarchical categories by the machine learning algorithm to facilitate subsequent structured output containing key information in the document. The framework essentially includes three parts: data pre-processing, feature engineering, and structural classification with a dynamic sample weighting algorithm.

In the data pre-processing part, we modify the Pdfplumber framework for PDF document information recognition and conversion and use JIEBA framework to carry out word segmentation. We convert the content into sentence-level text and text location information such as distance from page top and left. Then we remove the punctuations, the logos, and stop words in the text. The stop words are similar to 'the', 'is', 'at', 'which', 'on', etc. The stop words contain no insurance information and act as a sort of noise for structural classification. We generate a stop words list to identify and remove the stop words.

In the feature engineering part, we manually label the sentences according to the structure of the insurance document, such as level-1 heading, level-2 heading, content, etc. We also do context feature fusion which mainly incorporates features of the previous sentence into the current sentence features. See the detail of the context feature fusion in Section 2.2.

In the structural classification part, unlike the well-studied field of text semantic analysis and text classification, we carry out structural classification based on the sample features obtained in feature engineering. The sentences are classified into five structural categories: useless(-1), content(0), level-1 heading(1), level-2 heading(2), and level-3 heading(3). The data set details are shown in the Table 1. Through the classification, the content of the documents is segmented and divided into a lot of smaller parts. We store the content under the corresponding heading and save it as the dictionary data type. The key is the heading and the value is the corresponding content. With the key-value structure of the dictionary data type, we can extract the core information and output structural documents. (See output details in the Section 2.3).

In this structural classification part, the sample amount of each category is often quite different. The sample imbalance problem will thus occur. Even if the classification model is obtained, it is easy to over-rely on the limited data and lead to overfitting problems. When the model is applied to new data, the accuracy will not be high enough. To address this sample imbalance problem, dynamic weighting coefficients of each category are incorporated into the DWSA to conduct the classification task. We use accuracy and F1-score [41] to evaluate the model. The dynamic sample weighting algorithm adjusts the weight coefficients of each category. The weight coefficients of different categories are updated iteratively according to the classification F1-score and the feature importance. When the F1-score of the category with few samples is low, the algorithm will automatically assign the larger weight. Also, when the F1-score of other categories is affected and decreases, the algorithm will optimize the weight allocation to achieve the globally optimal result.

2. Approach

As shown in Figure 1, in this paper, we need to input a large number of unstructured Chinese insurance documents. The documents are in PDF format and are shown in appendix file Figure A1. The document content includes product name, title, insured liability, insured age, etc. Through the intelligent structural analysis, structured output of information is obtained. The specific steps are as follows:

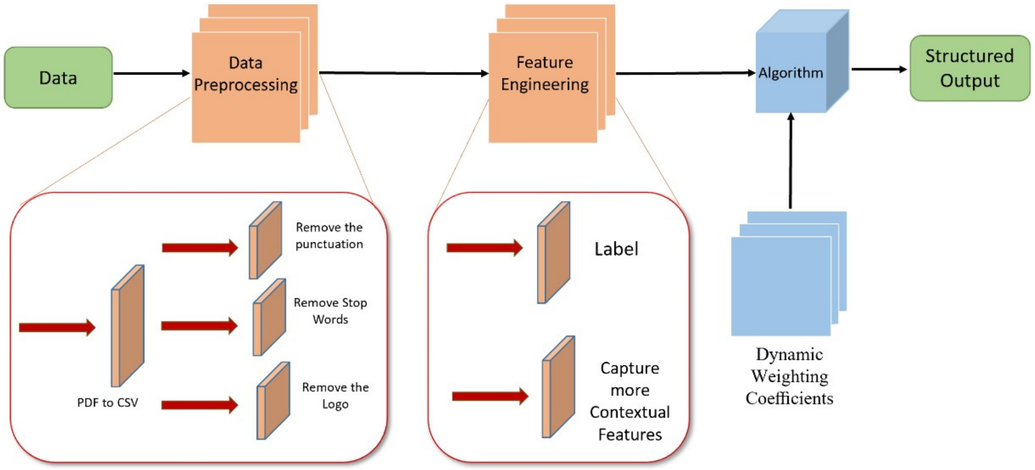


Figure 1. The framework of Dynamic Weighting Structural analysis (DWSA). The DWSA framework mainly includes three parts: data pre-processing, feature engineering, and structural classification.

2.1. Data Pre-Processing

The existing insurance documents are mostly saved in PDF format, and document content information needs to be converted before structural analysis operation. In this paper, we modify the Pdfplumber algorithm framework for PDF document contents conversion and convert the document content into sentence-level text. In terms of the treatment of broken sentences, a large number of sentences become incomplete sentences in the recognition and conversion step. The modified Pdfplumber framework can fit in this situation. For the insurance field, we do sentence segmentation based on semantic information. The broken sentences need to be treated differently in different circumstances.

Then, we do word segmentation and remove the punctuations and stop words in the text. There is no natural boundary between words in Chinese, so before the classification, the text should be divided into words. JIEBA uses a Chinese thesaurus to determine the probability of association between Chinese characters and can do word segmentation. When we use JIEBA to do Chinese word segmentation, all the words in the sentence are divided. The stop words contain no insurance information and constitute a sort of noise for structural classification. For stop words, we generate a stop words list and we use the list to identify the stop words and remove the stop words by programs.

The modified Pdfplumber framework is also used to extract document content features. The features include information such as font size, the font, coordinate position, and context spacing, etc. The extracted document feature information is further processed in depth in the data cleaning and feature engineering.

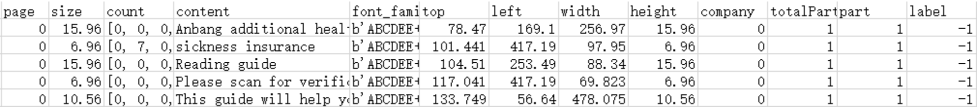
2.2. Feature Engineering

After data pre-processing, document content containing various feature information is obtained. Also, the Chinese word segmentation is carried out using the JIEBA framework. We label the key information such as headings at all levels and core attributes. The details of the labeled data set are shown in Table 1. After the conversion, the document data format is shown in Figure 2. The original content is in Chinese and we translate it into English as an alternative for better understanding. See the appendix file Figure A2 for the original Chinese content.

In addition to sentence-level content, we also get other feature information. These features are important for structural classification. For example, as shown in Figure 2, the document content belonging to the level-1 heading can be significantly different from contents of other levels in terms of font and font size features. Therefore, we can distinguish the level-1 headings by font size and other features.

Table 1. Details of the data set. The data set includes five categories and has 11,300 samples.

Class	Content	Number
Useless(-1)	Useless content in the document such as footer, special symbols, catalogue, etc.	2469
Content(0)	The main and useful content of the document	6704
Level-1 Heading(1)	The level-1 heading in the document	469
Level-2 Heading(2)	The level-2 heading in the document	1575
Level-3 Heading(3)	The level-3 heading in the document	83



page	size	count	content	font_fami	top	left	width	height	company	totalPart	part	label
0	15.96	[0, 0, 0,	Anbang additional heal	b'ABCDEE+	78.47	169.1	256.97	15.96	0	1	1	-1
0	6.96	[0, 7, 0,	sickness insurance	b'ABCDEE+	101.441	417.19	97.95	6.96	0	1	1	-1
0	15.96	[0, 0, 0,	Reading guide	b'ABCDEE+	104.51	253.49	88.34	15.96	0	1	1	-1
0	6.96	[0, 0, 0,	Please scan for verifi	b'ABCDEE+	117.041	417.19	69.823	6.96	0	1	1	-1
0	10.56	[0, 0, 0,	This guide will help y	b'ABCDEE+	133.749	56.64	478.075	10.56	0	1	1	-1

Figure 2. The file format in the data pre-processing step. We modify the Pdfplumber algorithm framework for PDF document information conversion, which can get more features in document content. The features mainly include size, font family, top, company, etc. The features named top, left, width, and height mean distance from page top, distance from page left, total width, and the total height of the sentence, respectively.

Also, context feature information can improve the performance of the structural classification. We correlate the feature information of individual words and sentences with multi-sentence and multi-word information. As shown in Figure 3, X_2 is the feature before the context feature fusion. X_2 will fuse with the feature of the previous sentence X_1 and generate the feature vector. Based on the existing features such as: left, top, height, etc., we obtain the context features named preleft, pretop, etc. which contain the content and location information of the previous sentence. The features of the previous sentence can provide additional useful information to the training sample. We vectorize the feature information and concatenate the context features and the current sentence features. Each sentence has many features. When we do vectorization, each feature counts as one dimension. Therefore, we use the context features to increase the dimension of the feature information. The context information can be supplemented and enhanced to make the machine learning algorithm learn and fit the data better.

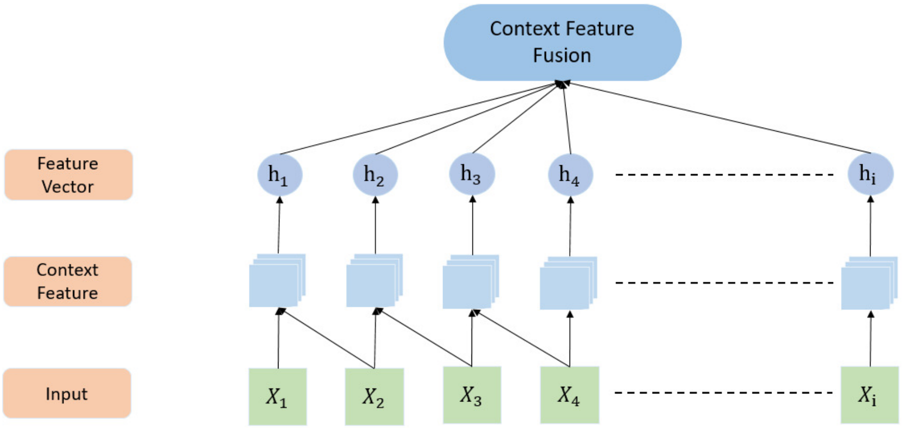


Figure 3. The context feature fusion algorithm. Through feature engineering, more context feature information is extracted. The context feature information is mapped into vectors for fusion.

2.3. Dynamic Weighting Algorithm

Unlike the well-studied field of text semantic analysis and text classification, in this part, based on the sample features obtained in feature engineering, we need to do sentence-level hierarchy classification for document content. In general, the insurance document will have three levels of headings. Also, we need to separate useless content from useful content in document content. Therefore, the sentences with features in documents are classified into five categories: useless(-1), content(0), level-1 heading(1), level-2 heading(2), and level-3 heading(3). Besides semantic information features, in feature engineering we also extract other features such as text coordinate information, font, font size, etc. Therefore, the language model is not effective for our task. We propose the Dynamic Weighting Algorithm derived from the Adaboost algorithm. Adaboost is an algorithm that can upgrade a weak learner to a strong learner. The working mechanism of the algorithm is as follows.

- An initial base learner is trained with the training set with equal weight of each sample.
- The sample weights of the training set are adjusted according to the predicted performance of the learner in the previous round. The algorithm increases the weight of the misclassified samples so that they will receive more attention in the next round of training. The base learner with a low error rate has a high weight, while the base learner with a high error rate has a low weight. Then train a new base learner.
- Repeat the second step until *M* base learners are obtained, and the final integration result is a combination of *M* base learners (*M* is the number of learners).

This structural classification task needs to divide document content information into 5 categories. The number of samples of each category is quite different. The details of the labeled data set are shown in Table 1. There are relatively more ‘content’ and ‘useless’ samples, while there are relatively fewer level-1, level-2, and level-3 heading samples. In the ordinary unweighted Adaboost algorithm, the weight of each category is the same and fixed. The F1-score of the category with fewer samples is lower. This has a significant impact on the subsequent structured output. If the title is not accurately classified, it is difficult to properly distinguish the document content.

To address this problem, we improve the classification algorithm and propose the Dynamic Weighting algorithm. The theoretical formula of the algorithm is as follows:

1. Initialize the sample weight D_1.

$$D_1 = (w_{11}, w_{12}, ...w_{1N}, w_{1i} = \frac{1}{N}, i = 1, ..., N) \tag{1}$$

where w_{1i} means the sample weight, N is the number of samples.

In the training task of this paper, there are five categories: useless content(-1), body content(0), level-1 heading(1), level-2 heading(2), level-3 heading(3), etc., and the initial category weight value is given by.

$$\alpha_{-1} = \alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 1 \tag{2}$$

where α means the category weight value.

2. Training base learner

For $m = 1, 2, ..., M$, repeat the following operations to get M base learners. We use the Decision Tree algorithm as the base learners in this task.

(1) Corresponding to the m_{th} sample weight D_m, the m_{th} base learner $G_m(x)$ was obtained.

(2) Calculate the classification error rate e_m of $G_m(x)$ on the weighted training data set.

$$e_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i) \tag{3}$$

where I means the event probability, $G_m(x_i)$ means the base learner, y_i means the labeled category.

(3) Calculate the coefficient of $G_m(x)$.

$$a_m = \frac{1}{2}\log\frac{1-\ e_m}{e_m} \tag{4}$$

where e_m is the error rate of $G_m(x)$, a_m is the weighting coefficient of $G_m(x)$, and a_m increases with the decrease of e_m.

(4) Update the weight of the training sample.

$$D_{m+1} = (w_{m+1,1}, w_{m+1,2}, \ldots, w_{m+1,N}) \tag{5}$$

$$w_{m+1,i} = \frac{w_{mi}}{Z_m}\exp\left(-a_m y_i G_m(x_i)\right) \tag{6}$$

where M means the number of learners, N means the number of samples. Z_m is a normalization factor, making the sum of the weights corresponding to all samples equal to 1.

(5) Update the weights of the categories.

$$\max(\sum_{j=-1}^{3}\varphi\Big(tag_{j,i}, \alpha_j\Big)) \tag{7}$$

$$j = -1, 0, 1, 2, 3; i = 1, 2, ..., N \tag{8}$$

where φ means the F1 score calculation function, $tag_{j,i}$ means the samples in different categories.

3. The final classifier.

$$G(X) = sign(\sum_{i=1}^{M} a_m G_m(x)) \tag{9}$$

where $sign$ is a function. When $x > 0, sign(x) = 1; x = 0, sign(x) = 0; x < 0, sign(x) = -1$.

According to Equation (4), when the error rate of $G_m(x)$ of the base learner is $e_m < 0.5$, $a_m > 0$, and a_m increases with the decrease of e_m. The proportion of the base learner with lower classification error rate is larger in the final integration. The algorithm model is able to adapt to the training error rate of each weak classifier.

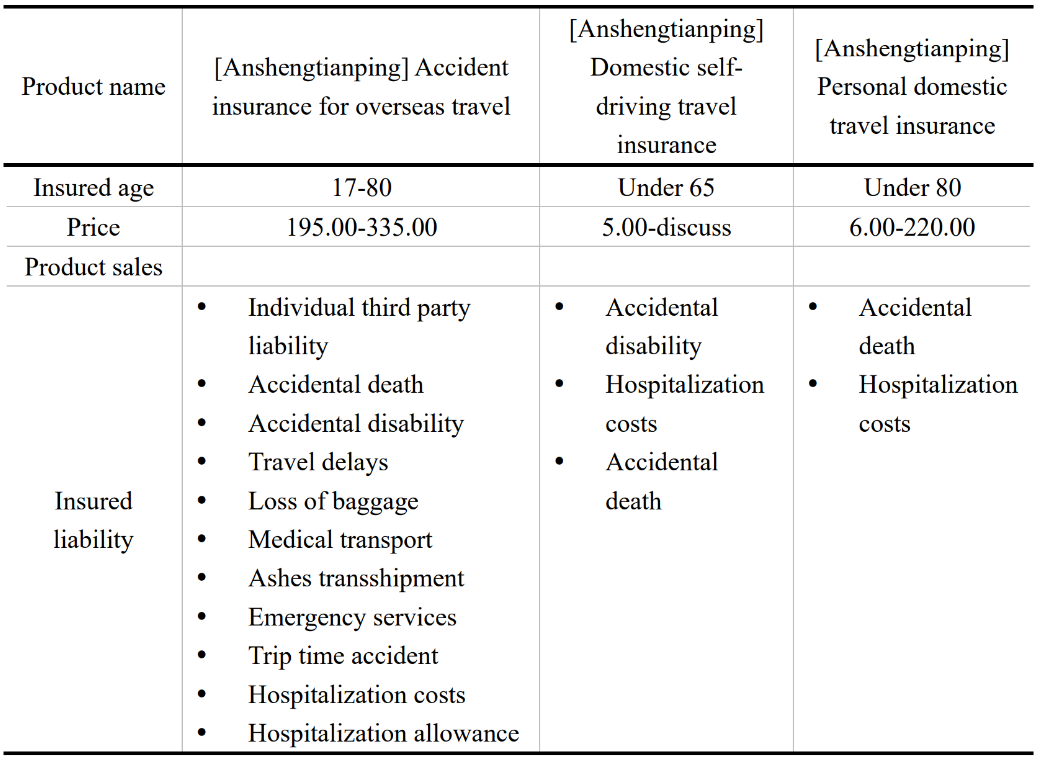
According to Equations (2) and (7), the weight is initialized first according to the number of various category tags. In the model training process, according to the iteration of the algorithm, the weight value is fine-adjusted with the learning rate γ, and we set a threshold β to achieve the optimal weight allocation and achieve the highest comprehensive F1 score. If $e_m > \beta$, $\alpha_j = \alpha_j + \gamma$. The computation complexity of dynamic weighting coefficients is $O(N)$. The flow chart of the algorithm is shown in Algorithm 1. The experimental results and weight coefficients are shown in Section 3.

As shown in Figure 4, through the classification, the content of documents is segmented and classified according to its hierarchy. The data processed by the algorithm model is reorganized to generate the output. We save the content under the corresponding heading and save it as the dictionary data type. The key is the heading and the value is the corresponding content. The headings at each level are nested. In content extraction, the content can be located quickly through the nested relation of headings at all levels. For example, we can see in Figure 4, 'Insured liability' is the level-3 heading saved in level-2 heading, and the contents belong to the heading are shown on the right. When several similar insurance documents need to be compared, DWSP can be used for structural comparison. The example shows the comparison result of three insurance documents.

Algorithm 1 The Dynamic Weighting Algorithm

```
**Input**:
TrainingSet $D = (x_1, y_1), (x_2, y_2), ...(x_n, y_n)$
Base Learner $G_m(x)$
Iterations $M$
**Begin**:
Initialize the sample weight and the category weight
$D_1(x) = \frac{1}{N}$
$\alpha_{-1} = \alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 1$
**for** $t = 1, 2, .....M$ **do**
    $G_m(x_i, y_i)$
    $e_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i)$
    **if** $e_m < 0.5$ **then**
        $a_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$
        Update the weight value
        $D_{m+1} = (w_{m+1,1}, w_{m+1,2}, \ldots w_{m+1,N})$
        $w_{m+1,i} = \frac{w_{mi}}{Z_m} \exp(-a_m y_i G_m(x_i))$
        **if** $e_m > \beta$ **then**
            $\alpha_j = \alpha_j + \gamma$
            $\max(\sum_{j=-1}^{3} \varphi\left(tag_{j,i}, \alpha_j\right))$
        **end if**
    **end if**
**end for**
$G(X) = sign(\sum_{i=1}^{M} a_m G_m(x))$

**Output**: Get the Dynamic Weighting Algorithm calculation result
```



Product name	[Anshengtianping] Accident insurance for overseas travel	[Anshengtianping] Domestic self-driving travel insurance	[Anshengtianping] Personal domestic travel insurance
Insured age	17-80	Under 65	Under 80
Price	195.00-335.00	5.00-discuss	6.00-220.00
Product sales			
Insured liability	• Individual third party liability • Accidental death • Accidental disability • Travel delays • Loss of baggage • Medical transport • Ashes transshipment • Emergency services • Trip time accident • Hospitalization costs • Hospitalization allowance	• Accidental disability • Hospitalization costs • Accidental death	• Accidental death • Hospitalization costs

Figure 4. Structured documents output. Document information is extracted for structured comparison. (The language of insurance documents is Chinese. We translate the structured output into English for better understanding. See the appendix file Figure A3 for the Chinese output).

3. Results and Discussion

In the experiments of this paper, 11,300 data samples are collected from insurance companies, and we use about 9040 samples (80%) as the training set, and use the 2260 samples (20%) as the test set. We first process the dataset through data pre-processing and feature engineering. Then, we conduct the classification algorithm comparison in experiments. The experiments in this paper are supervised learning tasks for machine learning and all samples are labeled.

3.1. Results of Some Existing Algorithms

The experimental samples are Chinese insurance document contents with features. The feature representation of each sample has strong attributes such as high dimension, nonlinearity, and data sparsity. For the classification task, the Support Vector Machine (SVM) algorithm, the Xgboost algorithm, the Lightgbm Algorithm, the TextCNN Algorithm, and the Adaboost algorithm are selected for comparison of classification accuracy and efficiency. The performance of each algorithm is evaluated by the comprehensive accuracy and the classification F1-score of each category. Finally, the results of the algorithms are compared.

Figure 5 shows the total accuracy of each algorithm. The accuracy of the Xgboost algorithm reaches 91.93% and the TextCNN algorithm achieves 91.56%. The TextCNN algorithm is mainly used to extract semantic information of text content, which is not effective in our structural classification task. In addition, due to the imbalance of samples, the Macro F1-score of the TextCNN algorithm is low. The SVM algorithm has lower accuracy of 89.88%. The accuracy of the Lightgbm algorithm is 92.28%, and that of the Adaboost algorithm is 91.32%. In the second stage of the experiment, we modify the Adaboost algorithm and add dynamic weight calculation, consequently the accuracy of the DWSA algorithm is increased to 94.68%. Also, the statistical comparison between the proposed method and other algorithms is conducted. As shown in Table 2, we calculate the P-value and do the Bonferroni correction. The significance level k is 0.05 and a total of 15 groups are compared in pairs. Therefore, after the Bonferroni correction, the significance level $k^{'} = k/15 = 0.0033$.

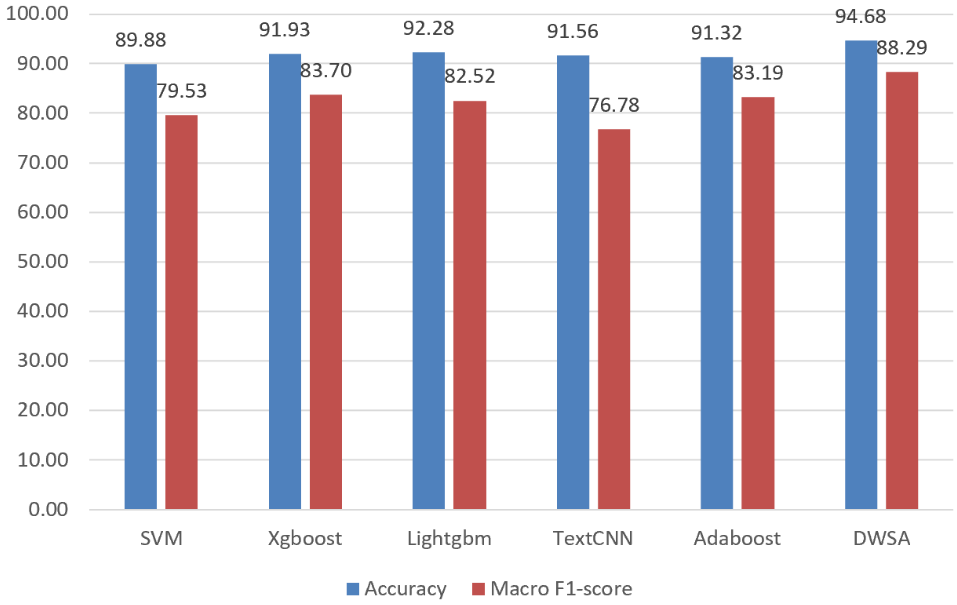


Figure 5. The result of the algorithm comparison experiment on the insurance data set. Accuracy and Macro F1-score are used to evaluate different algorithms. The proposed DWSA has the accuracy of 94.68% (3.36% absolute improvement over Adaboost) and the F1-score of 88.29% (5.1% absolute improvement over Adaboost).

Table 2. The statistical comparison between the proposed method and other algorithms. The significance level is 0.05. After the Bonferroni correction, the significance level is 0.0033.

Comparison	Bonferroni Correction	*P*-Value
SVM-DWSA	-	0.001
Xgboost-TextCNN	-	0.001
LightGBM-DWSA	-	0.001
TextCNN-Adaboost	-	0.001
TextCNN-DWSA	-	0.001
Adaboost-DWSA	-	0.001
LightGBM-TextCNN	-	0.002
Xgboost-DWSA	0.0033	0.003
SVM-Xgboost	-	0.015
SVM-Adaboost	-	0.034
SVM-LightGBM	-	0.087
SVM-TextCNN	-	0.144
Xgboost-LightGBM	-	0.474
LightGBM-Adaboost	-	0.678
Xgboost-Adaboost	-	0.764

To test the generalization ability of our proposed classifier in DWSA in other fields, we use the public data set Baike2018qa [42] and Iris [43] to conduct the comparison. Please note that unlike the well-studied field of text semantic analysis and text classification, published work on document structural analysis is still scarce and public data sets similar to insurance documents that contain rich hierarchical information are hard to find. The Baike2018qa contains Chinese question-and-answer texts in various fields, which can be used for classification at sentence level. However, the Baike2018qa data set only has semantic information without other features such as font size, the font, coordinate position, and context spacing, etc. We select a subset of the data set. There are five categories ('health', 'game', 'life', 'education', and 'entertainment') with unbalanced samples. The Iris data set contains 3 categories, with each sample containing 4 features, which can be used for classification. The result is shown in Figure 6. The obvious better performance on our collected insurance data set than that on the public data sets can be attributed to the effective data pre-processing and feature engineering in our framework.

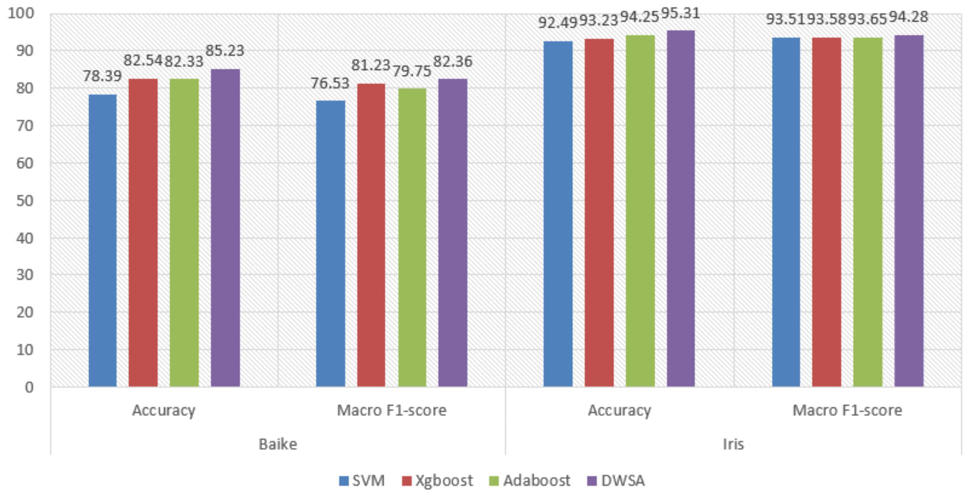


Figure 6. The result of the algorithm comparison experiment on the public data set Baike2018qa and Iris.

In addition to the standard of comprehensive accuracy, we also introduce the Macro F1-score and the classification F1-score of each category. In the experiment of this paper, there are more samples of body content(0) and useless content(-1), while there are fewer samples of level-1 heading(1), level-2 heading(2), and even fewer samples of level-3 heading(3). Therefore, in algorithm training, the comprehensive accuracy sometimes cannot reflect the real classification results well. We introduce the Macro F1-score and the classification F1-score of each category to better analyse and improve the performance of the algorithm. As shown in the Table 3, the performance of the algorithms is slightly different, but the body content(0) and the useless content(-1) are all better in the training of the algorithms due to a large amount of sample data. Among them, the Adaboost algorithm is 92% in the F1-score of these two categories. In terms of the F1-score of level-1 heading(1) and level-3 heading(3), the F1-score of the algorithms is lower than that of body content(0). Therefore, we improve it in the second stage experiment.

Table 3. Experimental results of insurance data set include precision, recall, and f1-score. The SVM algorithm, the Xgboost algorithm, the Lightgbm algorithm, the TextCNN algorithm, and the Adaboost algorithm are selected for comparison. The bold numbers are the best results.

Algorithm		Useless(-1)	Content(0)	Level-1 Heading(1)	Level-2 Heading(2)	Level-3 Heading(3)
SVM	Precision	0.91	0.93	0.65	0.82	0.75
	Recall	0.88	0.95	0.62	0.89	0.52
	F1-score	0.90	0.94	0.63	0.86	0.61
Xgboost	Precision	0.96	0.92	0.88	0.87	0.89
	Recall	0.86	0.96	0.69	0.95	0.52
	F1-score	0.91	0.94	0.77	0.91	0.66
Lightgbm	Precision	0.94	0.90	0.97	0.95	0.86
	Recall	0.90	0.98	0.58	0.91	0.47
	F1-score	0.92	0.94	0.73	0.93	0.61
TextCNN	Precision	0.90	0.92	0.63	0.87	0.80
	Recall	0.83	0.96	0.58	0.87	0.39
	F1-score	0.86	0.94	0.60	0.87	0.52
Adaboost	Precision	0.95	0.93	0.85	0.90	0.98
	Recall	0.91	0.92	0.62	0.89	0.53
	F1-score	0.92	0.92	0.72	0.89	0.69
DWSA	Precision	0.93	**0.96**	0.85	0.94	0.87
	Recall	**0.92**	0.97	**0.80**	**0.96**	**0.66**
	F1-score	**0.93**	**0.97**	**0.82**	**0.95**	**0.75**

3.2. The DWSA Experimental Result

In the second stage experiment of this paper, based on the previous results, the proposed DWSA model is used.

We have added the dynamic weight algorithm technique to the existing algorithm (the theory in Section 2). We give level-1 heading(1) and level-3 heading(3) the more weight. The calculated weight of each category: $\alpha_{-1} = 1.1, \alpha_0 = 1$, $\alpha_1 = 1.4, \alpha_2 = 1.2$, $\alpha_3 = 2.8$. In the second stage experiments, different weight values will be obtained according to different sample numbers and feature weights of each category. Based on the principle of iterative updating parameters with the algorithm, the weights of training samples were dynamically updated together with the weights of each classification category to obtain better training results. The learning rate γ is 0.1. The experimental results are shown in the Table 3 and Figure 7.

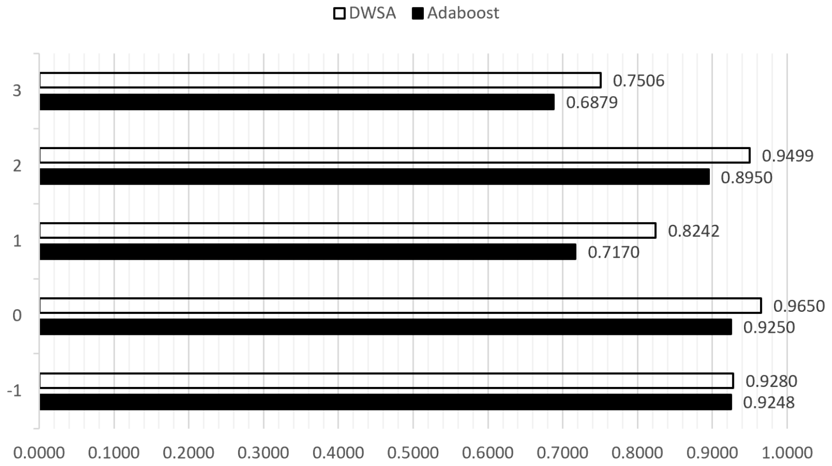
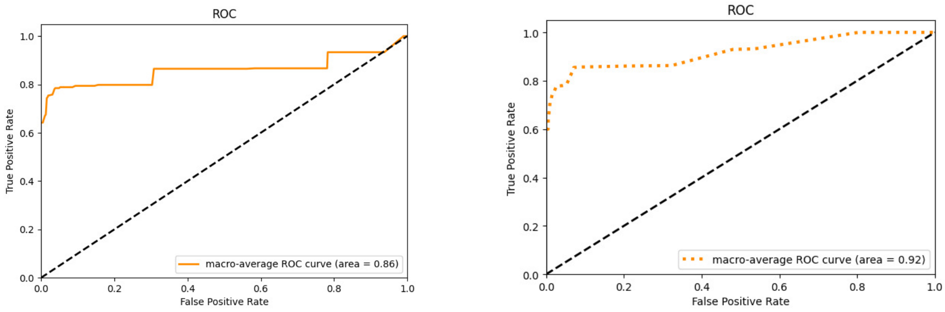


Figure 7. The comparison of the DWSA and the Adaboost algorithm (Insurance data set). The improved model increases the F1-score of the level-1 heading(1) from 71.7% to 82.42%. For the level-3 heading(3), the F1-score is increased by 6.27%. The accuracy of the level-2 heading(2) is improved by 5.49% and reaches 94.99%. The weights of useless content(-1) and body content(0) are reduced relatively, but the accuracy is still increased by 0.32% and 4%, and the final accuracy achieves 92.8% and 96.5%.

It can be seen that DWSA has improved significantly in terms of comprehensive accuracy and F1-score.

According to the data shown in the Figures 5 and 7, and Table 3, the accuracy of the DWSA algorithm is improved from 91.32% to 94.68%. In terms of the F1-score of each category, the level-1 heading(1) and the level-3 heading(3) have the most obvious improvements, with increases of 10.72% and 6.27% respectively. The F1-score of the level-2 heading(2) is improved by 5.49% because of the high base value (89.5%). The final F1-score achieves 94.99%. In the DWSA algorithm, the weights of useless content(-1) and body content(0) are reduced relatively, but the F1-score still increases by 0.32% and 4% relatively, and the final F1-score achieves 92.8% and 96.5%, respectively. The results show that according to the dynamic weight algorithm principle of iterative updating, the algorithm model can flexibly adjust the category weight. Even if the weight is reduced, the category with a large number of sample data will get a reasonable weight eventually and keep the F1-score stable or slightly improved. Therefore, the results still have good performance.

Also, as shown in Figures 8 and 9, we visualize the feature importance and plot the ROC (Receiver Operating Characteristic) curves of the Adaboost algorithm and the DWSA algorithm. By comparing the ROC curves, we can see that the improved DWSA algorithm has better performance. The DWSA algorithm has a larger AUC (Area Under Curve) and a smoother ROC curve.



(a) Adaboost (b) DWSA

Figure 8. The Macro-average ROC curve of the Adaboost algorithm and The DWSA algorithm. (Insurance data set) (**a**) Adaboost; (**b**) DWSA.

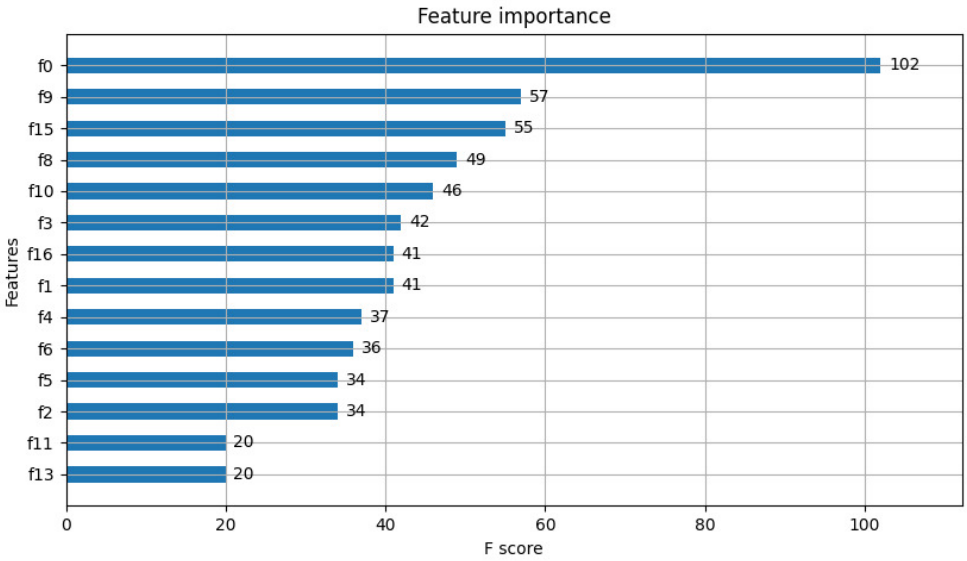


Figure 9. The feature importance of data. F0–F16 respectively represent the features, including size, count, content, font-family, top, left, width, page, height, company, pre-left, pre-size, part, pre-font, total-part, pre-top, and pre-behind. The feature importance of 'page', 'part', and 'total-part' is too low to show in the figure; 'size' means the font size, 'count' means the word and punctuation count, 'content' means the content in documents, 'font-family' means the font, 'company' means the insurance company, 'width' means the total width of the sentence, 'page' means the page number, 'height' means the total height of the sentence, 'part' means the order of the sentence in a paragraph, 'total-part' means the number of sentences in a paragraph, 'top' and 'left' mean the coordinate position of each sentence in the page, 'pre-left', 'pre-size', 'pre-font', and 'pre-top' are the corresponding features of the previous sentence. ('pre-left', 'pre-top' and so on are just the names we call the features). 'pre-behind' means the distance from page bottom of the previous sentence. The previous sentence features are incorporated into context feature fusion.

3.3. Experimental Summary

In this experiment, the algorithms commonly used in data mining and classification are compared and tested first.

In the second stage experiment, the DWSA algorithm is tested. By assigning the dynamic weight value, the performance of the algorithm is significantly improved in terms of the comprehensive accuracy and the classification F1-score of each category.

Finally, the data processed by the DWSA framework is reorganized and outputted as structured information.

4. Conclusions

In this paper, a framework named DWSA for intelligent structural analysis of Chinese insurance documents is proposed. We use the techniques of data pre-processing, feature engineering, and structural classification to process the documents. The proposed framework provides a convenient platform for insurance practitioners and related researchers to survey insurance documents. Also, the DWSA can effectively address the problem of sample imbalance by the dynamic sample weighting algorithm. Verified by experiments, the DWSA algorithm can significantly improve the comprehensive accuracy and the classification F1-score of each hierarchical category. The comprehensive accuracy reaches 94.68% (3.36% absolute improvement) and the Macro F1-score achieves 88.29% (5.1% absolute improvement).

Author Contributions: Conceptualization, T.Y., Y.L. and Z.H.; Data curation, T.Y.; Formal analysis, T.Y. and Z.H.; Investigation, T.Y.; Methodology, T.Y.; Project administration, Z.H.; Resources, Y.L. and Z.H.; Software, T.Y.; Supervision, Y.L. and Z.H.; Validation, T.Y.; Visualization, T.Y.; Writing—original draft, T.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The work described in this paper was supported by the National Key R&D Program of China (2017YFB0403602 and 2016Y-FB0400603), the National Natural Science Foundation of China (Grant 11074005 and 91021007), and the BUPT innovation and entrepreneurship support program for financial supports.

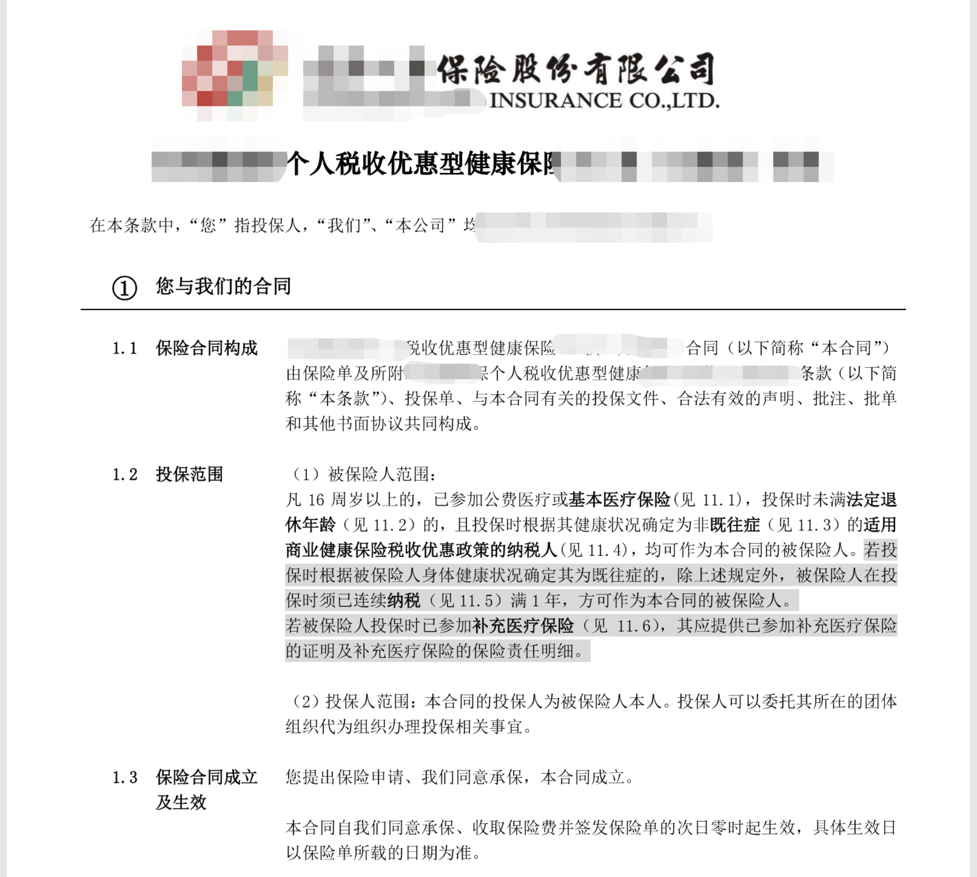
Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DWSA	Dynamic Weighting Structural Analysis
SVM	Support Vector Machines
ROC	Receiver Operating Characteristic
AUC	Area Under Curve

Appendix A



[redacted] 保险股份有限公司
INSURANCE CO.,LTD.

[redacted]个人税收优惠型健康保[redacted]

在本条款中，“您”指投保人，“我们”、“本公司”均[redacted]

① 您与我们的合同

1.1 **保险合同构成** [redacted]税收优惠型健康保险[redacted]合同（以下简称“本合同”）由保险单及所附[redacted]保个人税收优惠型健康[redacted]条款（以下简称“本条款”）、投保单、与本合同有关的投保文件、合法有效的声明、批注、批单和其他书面协议共同构成。

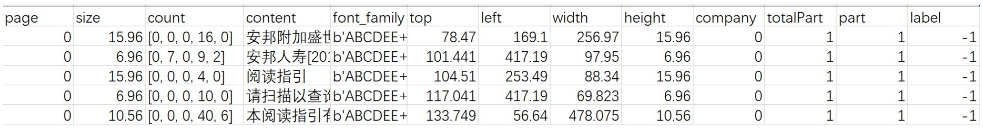
1.2 **投保范围** （1）被保险人范围：
凡 16 周岁以上的，已参加公费医疗或**基本医疗保险**(见 11.1)，投保时未满**法定退休年龄**（见 11.2）的，且投保时根据其健康状况确定为非**既往症**（见 11.3）的**适用商业健康保险税收优惠政策的纳税人**(见 11.4)，均可作为本合同的被保险人。若投保时根据被保险人身体健康状况确定其为既往症的，除上述规定外，被保险人在投保时须已连续**纳税**（见 11.5）满 1 年，方可作为本合同的被保险人。
若被保险人投保时已参加**补充医疗保险**（见 11.6），其应提供已参加补充医疗保险的证明及补充医疗保险的保险责任明细。

（2）投保人范围：本合同的投保人为被保险人本人。投保人可以委托其所在的团体组织代为组织办理投保相关事宜。

1.3 **保险合同成立及生效** 您提出保险申请、我们同意承保，本合同成立。

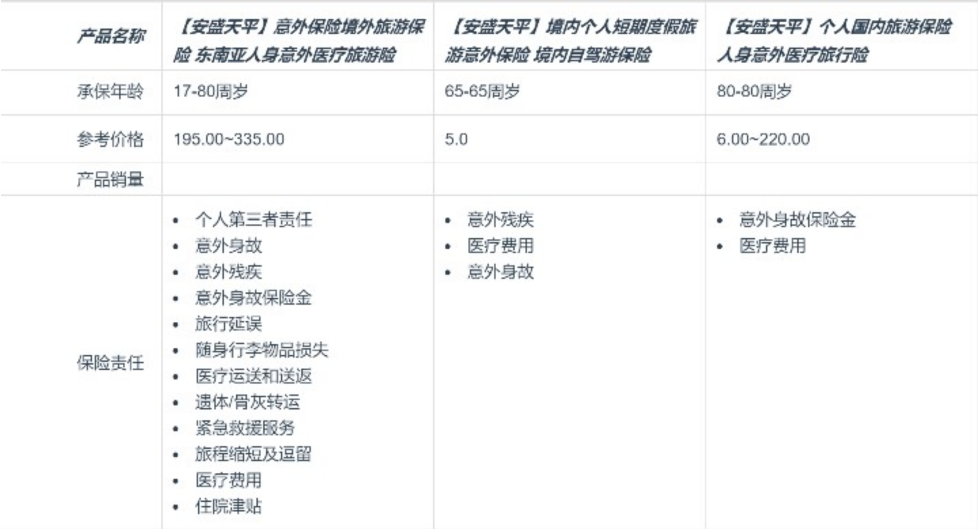
本合同自我们同意承保、收取保险费并签发保险单的次日零时起生效，具体生效日以保险单所载的日期为准。

Figure A1. A typical original Chinese insurance document in PDF format.



page	size	count	content	font_family	top	left	width	height	company	totalPart	part	label
0	15.96	[0, 0, 0, 16, 0]	安邦附加盛	b'ABCDEE+	78.47	169.1	256.97	15.96	0	1	1	-1
0	6.96	[0, 7, 0, 9, 2]	安邦人寿[20	b'ABCDEE+	101.441	417.19	97.95	6.96	0	1	1	-1
0	15.96	[0, 0, 0, 4, 0]	阅读指引	b'ABCDEE+	104.51	253.49	88.34	15.96	0	1	1	-1
0	6.96	[0, 0, 0, 10, 0]	请扫描以查	b'ABCDEE+	117.041	417.19	69.823	6.96	0	1	1	-1
0	10.56	[0, 0, 0, 40, 6]	本阅读指引	b'ABCDEE+	133.749	56.64	478.075	10.56	0	1	1	-1

Figure A2. The file format in the data pre-processing step (in Chinese). We modify the Pdfplumber algorithm framework for PDF document information conversion, which can get more features in documents content. The features mainly include size, font family, top, and company etc. The features named top, left, width, and height mean the context relative location information.



产品名称	*【安盛天平】意外保险境外旅游保险 东南亚人身意外医疗旅游险*	*【安盛天平】境内个人短期度假旅游意外保险 境内自驾游保险*	*【安盛天平】个人国内旅游保险人身意外医疗旅行险*
承保年龄	17-80周岁	65-65周岁	80-80周岁
参考价格	195.00~335.00	5.0	6.00~220.00
产品销量			
保险责任	• 个人第三者责任 • 意外身故 • 意外残疾 • 意外身故保险金 • 旅行延误 • 随身行李物品损失 • 医疗运送和送返 • 遗体/骨灰转运 • 紧急救援服务 • 旅程缩短及逗留 • 医疗费用 • 住院津贴	• 意外残疾 • 医疗费用 • 意外身故	• 意外身故保险金 • 医疗费用

Figure A3. Structured document output. Document information is extracted for structural comparison.(Chinese output)

References

1. Park, S.; Lee, W.; Lee, J. Learning of indiscriminate distributions of document embeddings for domain adaptation. *Intell. Data Anal.* **2019**, *23*, 779–797. [CrossRef]
2. Aggarwal, C.C.; Zhai, C.X. A survey of text classification algorithms. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 163–222.
3. Neetu, Hierarchical classification of web content using naive bayes approach. *Int. J. Comput. Sci. Eng.* **2013**, *5*, 402–408. http://www.enggjournals.com/ijcse/doc/IJCSE13-05-05-117.pdf (accessed on 5 May 2013).
4. Thirunavukkarasu, K.S.; Sugumaran, S. Analysis of classification techniques in data mining. *Int. J. Eng. Sci. Res. Technol.* **2013**, *2*, 779–797.
5. Peng, N.; Zhou, X.; Niu, B.; Feng, Y. Predicting Fundraising Performance in Medical Crowdfunding Campaigns Using Machine Learning. *Electronics* **2021**, *10*, 143. [CrossRef]
6. Chakroun, I.; Aa, T.V.; Ashby, T.J. Guidelines for enhancing data locality in selected machine learning algorithms. *Intell. Data Anal.* **2019**, *23*, 1003–1020. [CrossRef]
7. Du, X.; Cardie, C. Document-level event role filler extraction using multi-granularity contextualized encoding. *arXiv* **2020**, arXiv:2005.06579.
8. Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; Chang, S.F. Cross-media structured common space for multimedia event extraction. *arXiv* **2020**, arXiv:2005.02472.
9. Wu, J.; Cai, Z.; Zeng, S.; Zhu, X. Artificial immune system for attribute weighted naive bayes classification. In Proceedings of The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–8.
10. Zhao, J.; Forouraghi, B. An interactive and personalized cloud-based virtual learning system to teach computer science. In *Advances in Web-Based Learning—ICWL 2013*; Wang, J.F., Lau, R., Eds.; Springer: Berlin, Germany, 2013.
11. Khan, A.; Ilyas, T.; Umraiz, M.; Mannan, Z.I.; Kim, H. CED-Net: Crops and Weeds Segmentation for Smart Farming Using a Small Cascaded Encoder-Decoder Architecture. *Electronics* **2020**, *9*, 1602. [CrossRef]
12. Zhang, S.; Wu, G.; Gu, J.; Han, J. Pruning Convolutional Neural Networks with an Attention Mechanism for Remote Sensing Image Classification. *Electronics* **2020**, *9*, 1209. [CrossRef]
13. Kravcik, M.; Wan, J. Towards open corpus adaptive e-learning systems on the web. In *Advances in Web-Based Learning—ICWL 2013*; Wang, J.F., Lau, R., Eds.; Springer: Berlin, Germany, 2013.
14. Lu, H.-Y.; Kang, N.; Li, Y.; Zhan, Q.-Y.; Xie, J.-Y.; Wang, C.-J. Utilizing Recurrent Neural Network for topic discovery in short text scenarios. *Intell. Data Anal.* **2019**, *23*, 259–277. [CrossRef]

15. Zhang, Z.; Lin, H.; Li, P.; Wang, H.; Lu, D. Improving semi-supervised text classification by using Wikipedia knowledge, In *Web-Age Information Management*; Wang, J., Xiong, H., Ishikawa, Y., Xu, J., Zhou, J., Eds.; Springer: Berlin, Germany, 2013.
16. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
17. Dhaliwal, S.S.; Nahid, A.A.; Abbas, R. Effective intrusion detection system using XGBoost. *Information* **2018**, *9*, 149. [CrossRef]
18. Ogunleye, A.; Wang, Q.G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *17*, 2131–2140. [CrossRef] [PubMed]
19. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
20. Gómez, P.; Roncancio, C.; Casallas, R. Analysis and evaluation of document-oriented structures. *Data Knowl. Eng.* **2021**, *134*, 101893. [CrossRef]
21. Liu, L.; Wang, Z.; Qiu, T.; Chen, Q.; Lu, Y.; Suen, C.Y. Document image classification: Progress over two decades. *Neurocomputing* **2021**, *453*, 223–240. [CrossRef]
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O. Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2008**, *16*, 321–357. [CrossRef]
23. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 318–327. [CrossRef]
24. Han, X.; Zhao, J. Named entity disambiguation by leveraging wikipedia semantic knowledge. In Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, 2–6 November 2009; pp. 215–224.
25. Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; Li, J. A unified mrc framework for named entity recognition. *arXiv* **2019**, arXiv:1910.11476.
26. Pennacchiotti, M.; Pantel, P.; Entity extraction via ensemble semantics. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Singapore, 6–7 August 2009; pp. 238–247.
27. Rijhwani, S.; Preoţiuc-Pietro, D. Temporally-Informed Analysis of Named Entity Recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online Event, 6–8 July 2020; pp. 7605–7617.
28. Sikelis, K.; Tsekouras, G.E.; Kotis, K. Ontology-Based Feature Selection: A Survey. *Future Internet* **2021**, *13*, 158. [CrossRef]
29. Han, X.; Zhao, J. Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 11–16 July 2010; pp. 50–59.
30. Milne, D.; Witten, I.H. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley, CA, USA, 26–30 October 2008; pp. 509–518.
31. Shen, W.; Wang, J.; Luo, P.; Wang, M. Linking named entities in tweets with knowledge base via user interest modeling. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, IL, USA, 11–14 August 2013; pp. 68–76.
32. Zhang, W.; Sim, Y.C.; Su, J.; Tan, C.L. Entity linking with effective acronym expansion, instance selection and topic modeling. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 19–22 July 2011; pp. 1909–1914.
33. He, Z.; Liu, S.; Li, M.; Zhou, M.; Zhang, L.; Wang, H. Learning Entity Representation for Entity Disambiguation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 30–34.
34. Kulkarni, S.; Singh, A.; Ramakrishnan, G.; Chakrabarti, S. Collective annotation of Wikipedia entities in web text. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 457–466.
35. Yu, D.; Sun, K.; Cardie, C.; Yu, D. Dialogue-based relation extraction. *arXiv* **2020**, arXiv:2004.08056.
36. Zhang, T.; Liu, K.; Zhao, J. Cross Lingual Entity Linking with Bilingual Topic Model. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
37. Shi, C.; Quan, J.; Li, M. Information extraction for computer science academic rankings system. In Proceedings of the 2013 International Conference on Cloud and Service Computing, Beijing, China, 4–6 November 2013; pp. 69–76.
38. Li, X.; Yan, H.; Qiu, X.; Huang, X. Flat: Chinese ner using flat-lattice transformer. *arXiv* **2020**, arXiv:2004.11795.
39. Chen, M.; Jin, X.; Shen, D. Short text classification improved by learning multigranularity topics. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 19–22 July 2011; pp. 1776–1781.
40. Wang, K.; Zong, C.; Su, K.Y. A character-based joint model for chinese word segmentation. In Proceedings of the 23rd International Conferenceon Computational Linguistics, Beijing, China, 23–27 August 2010; pp. 1173–1181;
41. Elzeki, O.M.; Alrahmawy, M.F. Elmougy, S. A New Hybrid Genetic and Information Gain Algorithm for Imputing Missing Values in Cancer Genes Datasets. *Int. J. Intell. Syst. Appl.* **2019**, *11*, 20. [CrossRef]
42. NLP Chinese Corpus. Available online: https://drive.google.com/file/d/1_vgGQZpfSxN_Ng9iTAvE7hM3Z7NVwXP2/view (accessed on 5 July 2021).
43. Iris Data Set. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris (accessed on 5 July 2021).