Lecture Notes on Data Engineering
and Communications Technologies 155

Bernard J. Jansen
Qingyuan Zhou
Jun Ye *Editors*

Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)

Volume 1

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 155

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

Bernard J. Jansen · Qingyuan Zhou · Jun Ye
Editors

Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)

Volume 1

Springer

Editors
Bernard J. Jansen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar

Jun Ye
School of Computer Science
and Cyberspace Security
Hainan University
Haikou, Hainan, China

Qingyuan Zhou
School of Economics and Management
Changzhou Institute of Mechatronic
Technology
Changzhou, China

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-19-9372-5 ISBN 978-981-19-9373-2 (eBook)
https://doi.org/10.1007/978-981-19-9373-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Conference Committee

General Chair

Lin Shen, Changzhou Vocational Institute of Mechatronic Technology, China

Local Organizing Chairs

Jin Lou, Changzhou Vocational Institute of Mechatronic Technology, China
Weijie Gu, Changzhou Vocational Institute of Mechatronic Technology, China

Program Chairs

Bernard J. Jansen, Qatar Computing Research Institute, Qatar
Qingyuan Zhou, Changzhou Vocational Institute of Mechatronic Technology, China
Jun Ye, Hainan University, China

Publicity Chairs

Haibo Liang, Southwest Petroleum University, China
Sulin Pang, Jinan University, China

Program Committee

Ameer Al-Nemrat, University of East London, UK
Robert Ching-Hsien Hsu, Chung Hua University, China
Neil Yen, University of Aizu, Japan
Meng Yu, The University of Texas at San Antonio, USA
Shunxiang Zhang, Anhui University of Science and Technology, China
William Liu, Auckland University of Technology, NZ
Mustafa Mat Deris, Universiti Tun Hussein Onn Malaysia, Malaysia
Zaher AL Aghbari, Sharjah University, UAE
Guangli Zhu, Anhui University of Science and Technology, China
Raja Al Jaljouli, College of Computer Science and Engineering, Kingdom of Saudi Arabia
Abdul Basit Darem, University of Mysore, India
Vijay Kumar, VIT, India
Xiangfeng Luo, Shanghai University, China
Jemal Abawajy, Deakin University
Ahmed Mohamed Khedr, University of Sharjah, UAE
Xiao Wei, Shanghai University, China
Sabu M. Thampi, Indian Institute of Information Technology and Management, India
Huan Du, Shanghai University, China
Shamsul Huda, Deakin University, Australia
Zhiguo Yan, Fudan University, China
Rick Church, UC Santa Barbara, USA
Tom Cova, University of Utah, USA
Susan Cutter, University of South Carolina, USA
Yi Liu, Tsinghua University, China
Kuien Liu, Pivotal Inc., USA
Wei Xu, Renmin University of China, China
V. Vijayakumar, Professor and Associate Dean, SCSE, VIT Chennai, India
Abdullah Azfar, KPMG Sydney, Australia
Florin Pop, University Politehnica of Bucharest, Romania
Kim-Kwang Raymond Choo, The University of Texas at San Antonio, USA
Mohammed Atiquzzaman, University of Oklahoma, USA
Rafiqul Islam, Charles Sturt University, Australia
Morshed Chowdhury, Deakin University, Australia

Preface

Cognition is emerging as a new and promising methodology with the development of cognitive-inspired computing, cognitive-inspired interaction, and systems, which have the potential to enable a large class of applications. These applications have emerged with great potential to change our lives. However, recent advances in artificial intelligence (AI), fog computing, big data, and cognitive computational theory show that multi-disciplinary cognitive-inspired computing still struggles with fundamental, long-standing problems, such as computational models and decision-making mechanisms based on the neurobiological processes of the brain, cognitive sciences, and psychology. How to enhance human cognitive performance with machine learning, common sense, natural language processing, etc., are worth exploring.

The 2nd International Conference on Cognitive-based Information Processing and Applications (CIPA 2022) held in Changzhou, China, from September 22 to 23, 2022. The conference communicated the theory, technology, and application of artificial intelligence, including precision mining, intelligent computing, deep learning, and all other theories, models, and technologies related to artificial intelligence.

The purpose of CIPA2022 is to provide a forum for the presentation and discussion of innovative ideas, cutting-edge research results, and novel techniques, methods, and applications on all aspects of technology and intelligence in intelligent computing.

At least two independent experts reviewed each paper. The conference would not have been a reality without the contributions of the authors. We sincerely thank all the authors for their valuable contributions. We want to express our appreciation to all members of the Program Committee for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We want to express our thanks for the strong support of the publication chairs, organizing chairs, program committee members, and all volunteers.

Our special thanks are also due to the editor of Springer, Ramesh Nath Premnath, for his assistance throughout the publication process.

Doha, Qatar — Bernard J. Jansen
Changzhou, China — Qingyuan Zhou
Haikou, China — Jun Ye

Keynotes

Bernard J. Jansen is a principal scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is the editor-in-chief of the journal, *Information Processing and Management* (Elsevier), a member of the editorial boards of seven international journals, and the former editor-in-chief of the journal, *Internet Research* (Emerald). He has received several awards and honors, including an ACM Research Award, six application development awards, and a university-level teaching award, along with other writing, publishing, research, teaching, and leadership honors. Dr. Jansen has authored or co-authored 300 or so research publications, with articles appearing in a multi-disciplinary range of journals and conferences. He is the author of the book, *Understanding Sponsored Search: A Coverage of the Core Elements of Keyword Advertising* (Cambridge University Press).

Jemal Abawajy is a faculty member at Deakin University and has published more than 100 articles in refereed journals and conferences as well as a number of technical reports. He is on the editorial board of several international journals and edited several international journals and conference proceedings. He has also been a member of the organizing committee for over 60 international conferences and workshops, serving in various capacities, including the best paper award chair, general co-chair, publication chair, vice-chair, and program committee. He is actively involved in funded research in building secure, efficient, and reliable infrastructures for large-scale distributed systems. Toward this vision, he is working in several areas, including pervasive and networked systems (mobile, wireless network, sensor networks, grid, cluster, and P2P), e-science and e-business technologies and applications, and performance analysis and evaluation.

Contents

Cognitive-Inspired Computing Fundamentals and Computing Systems

Cognitive-Inspired Computing with Big Data

AI-Assisted Cognitive Computing Approaches

Internet of Cognitive Things

About the Editors

Bernard J. Jansen is a Principal Scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is editor-in-chief of the journal, *Information Processing and Management* (Elsevier), a member of the editorial boards of seven international journals, and former editor-in-chief of the journal, *Internet Research* (Emerald). He has received several awards and honors, including an ACM Research Award, six application development awards, and a university-level teaching award, along with other writing, publishing, research, teaching, and leadership honors. Dr. Jansen has authored or co-authored 300 or so research publications, with articles appearing in a multi-disciplinary range of journals and conferences. He is author of the book, *Understanding Sponsored Search: A Coverage of the Core Elements of Keyword Advertising* (Cambridge University Press).

Qingyuan Zhou received Ph.D. degrees from Sichuan University (SCU), Chengdu in 2014, respectively. Currently, he is a professor in Changzhou Institute of Mechatronic Technology, China. He has joined and accomplished five national and six provincial research programs. He has also authored/coauthored over 60 papers in international/national journals and conferences including *Computers in Human Behavior*, *Electronic Commerce Research*, *Intelligent Automation and Soft Computing* etc. His current research interests include electronic commerce, artificial intelligence, and computational economics.

Jun Ye received his B.S. degree in Applied Mathematics at Chongqing University. M.S. degree in Cryptography at Guilin University of Electronic Technology. Ph.D. in Xidian University, He is a high level talent of Hainan Province, and he is working at school of Computer Science and Cyberspace Security of Hainan University. His current research interests include computer science and information security. He has authored or co-authored more than 20 high level publications, and he is also a reviewer of many well-known journals. He is one of the high level talents of Hainan Province, and got the “First prize” of science and technology progress prize of Hainan Province in 2019.

Cognitive-Inspired Computing Fundamentals and Computing Systems

Bayesian Classification Algorithm in Recognition of Insurance Tax Documents

Meiying Jin

Abstract With the development of Internet technology and the country's vigorous promotion of the openness and digitization of government information, more and more official document information is published on government websites, and the construction of a digital government is imminent. The purpose of this paper is to study the identification of insurance tax documents based on Bayesian classification algorithm. This paper introduces the main structure of the insurance tax document classifier and the implemented system modules. Aiming at the limitation of Naive Bayes algorithm, the introduction of weighting factor is proposed to improve the classification accuracy. At the same time, considering the classification degree of domain vocabulary, the characteristic words in the insurance tax official document information dictionary will achieve a better and better distinguishing effect, and the parameter factor is introduced to increase the proportion. It can be seen from the experiment that the average accuracy of the improved weighted Bayesian identification is 0.95, and the improved classification algorithm can meet the identification needs of insurance tax documents.

Keywords Bayesian classification · Insurance taxation · Document recognition · Data mining

1 Introduction

Official documents, short for official documents, are documents with legal authority, standard formats used in the management of all social activities of human beings and state management [1]. As the most important carrier for expressing the will of the state, implementing government documents and regulations, regulating administrative law enforcement, and transmitting important information, to a certain extent,

M. Jin (✉)
School of Accounting, Dongying Vocational Institute, Dongying 257091, Shandong, China
e-mail: dyzyxykjxyjmy@163.com

B. J. Jansen et al. (eds.), *Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022)*,
Lecture Notes on Data Engineering and Communications Technologies 155,
https://doi.org/10.1007/978-981-19-9373-2_1

official documents are the continuation and supplement of national government documents and regulations [2, 3]. At present, the number of corresponding government official documents increases exponentially every year. The attributes of personnel, organizations, and institutions in all official documents need to be systematically analyzed and managed. It is very important to identify and extract corresponding objects from official documents, and it needs to be able to machine automatic identification and processing, so the entity identification of government documents becomes very important, and it plays a basic technical support for later extraction, search, correlation comparison, etc., which has great practical significance for social management [4, 5].

According to its probability model analysis, Bayes' theorem can well analyze and make inferences on text classification [6]. Hashmi provides a comprehensive analysis of modern methods utilizing deep neural networks, studying deep learning methods for table detection and table structure recognition. Furthermore, it provides a comprehensive overview of current state of the art and related challenges for table understanding in document images. Leading datasets and their complexity have been detailed along with quantitative results. Furthermore, a brief overview of promising directions in which table analysis in document images can be further improved is briefly outlined [7]. Sombra presents a Bayesian classification algorithm application capable of classifying lamb carcasses using two output categories (constructed and done). From various parameters collected from animal carcass measurements, two classifications were identified: one for construction and one for completion [8]. The Bayesian classification algorithm can effectively improve the recognition accuracy of the special text of official documents, and will further enhance the construction of intelligent government, improve the efficiency of government affairs, and improve the image of the government [9].

After a thorough analysis of text classification, this paper uses Naive Bayesian classification to design and implement an English text classification system. In this classification system, the conditional probability table of its feature items under this class is established for each class in the training stage, thereby greatly improving the speed of text classification in the classification stage. Finally, this paper uses the $F1$ test value and other indicators to evaluate the classification results of the classification system. The experiments show that the classification system has high recognition accuracy under the experimental data.

2 Research on Bayesian Classification Algorithm in Recognition of Insurance Tax Documents

2.1 Features of Government Official Documents

Government documents are usually used as official documents of the government and other related departments and have been widely used in the political field of our country. As a general official document, it needs to meet the standardized format requirements. The specific content involved includes information such as confidentiality period, text number, issuer, urgency, attachment description, release authority, and release date [10].

The computerization of government documents is an inevitable trend of social development: it greatly reduces the workload of government employees by automatically reviewing and creating documents of various official documents and automatically supplementing the information structure [11], and can effectively improve the effectiveness of social governance and national management. Recognition of government documents is one of the most basic tasks of government information automation, event extraction, and relationship extraction, and it can be used as the basis for tasks such as event extraction and relationship extraction. The research on the identification of government documents is a huge part of the whole system [12, 13].

2.2 Naive Bayesian Classification Method

Among the many Bayesian classification methods, the Naive Bayesian classification method is fast and accurate in classification, only needs to scan the data once, and has strong anti-interference and self-correction capabilities, so it has received more and more attention [14]. Naive Bayesian classification method is based on the assumption of conditional independence of attribute classes, that is, after a given category node, each attribute node representing the text is independent of each other.

Naive Bayesian classification method has been successfully applied in many classification fields and engineering practice due to the advantages of simple logic, stable algorithm, small time and space overhead, and no preference for specific datasets [15, 16]. When the class condition independence assumption holds, Naive Bayes classification method is the best classification method compared with other classification methods. However, in many practical situations, the Naive assumption of class condition independence cannot be established, that is, there is a strong correlation between document features, which reduces the classification accuracy of the Naive Bayesian classification method. But even so, the Naive Bayes classification method can still achieve better classification results under normal circumstances.

2.3 Eigenvector Weight Calculation

Feature item extraction refers to selecting the segmented words with information representing the content of the article from the word segmentation result set after the end of word segmentation. According to the Chinese word segmentation and feature extraction described above, the original email content is output as a feature vector, which roughly covers all the content of the email. Therefore, it is very important to be able to extract the characteristic attributes with representative lines as the word segmentation of sample emails [17, 18]. The methods of feature word selection are as follows:

(1) Use mathematical mapping and other methods to simplify the feature data to be extracted into fewer representative words;
(2) Some representative words can be directly extracted from the data to be processed, and some irrelevant data can be removed;
(3) Select some representative words according to experts in different fields;
(4) Use the relevant mathematical calculation method to select and find some data with high data calculation value. This method is more reliable because it is calculated through calculation. Because of the reduction of human interference, it is especially suitable for text classification and text mining research in the field. The most commonly used method for feature word extraction is to calculate the score value of all the separated feature words through a calculation function, set a threshold value, and select the feature word if the calculated value exceeds the threshold value.

3 Investigation and Research of Bayesian Classification Algorithm in Recognition of Insurance Tax Documents

3.1 Experimental Setup

The model of the machine used in the experiment is ASUS laptop, the main configuration of the machine is Intel(R) Core(TM) i7-4700HQCPU, the memory is 4G, the JAVA programming language is used, and it is completed on the Myeclipse8.6 development platform. After preprocessing, text representation, and feature selection, the Naive Bayes algorithm is first used to conduct experiments, and then the effect of the improved classifier is compared and analyzed.

For the development of an experiment, it is indispensable to set the evaluation index of the evaluation experiment or prediction. This paper uses recall rate, accuracy rate, and $F1$ value, which are the evaluation criteria and reference data that can most intuitively reflect the classification results of the data model. It is the result of the training model formed according to the training samples in the process of the experiment according to the classification principle of the algorithm. These values will fluctuate because the dataset is single, the training is insufficient, or the detection environment used by the classification model is not suitable, but they can still intuitively reflect the results of experiments and the predictions of multi-sample classification models.

3.2 Corpus Construction

The experiment used in this paper is a corpus of 2000 government official documents in a province. They are insurance tax documents, and non-insurance tax documents, of which 1051 are insurance tax documents and 949 non-insurance tax documents. In this paper, the dataset is divided into training set and test set according to the ratio of 8:2. After the division, the text data of the training set is 1600, including 840 texts of insurance tax documents and 760 texts of non-insurance tax documents; the text data of the test set are 400 articles. There are 400 articles, of which 211 are insured tax official documents and 189 are non-insured tax official documents.

3.3 Introduction of Weighting Factor

We introduce the parameter to represent the degree of discrimination of the category, and is defined as follows:

$$\alpha_t = \frac{\mathrm{TF}\left(w_t|C_j\right)}{\mathrm{TF}(w_t|C) + 0.1} \tag{1}$$

In the formula, $\mathrm{TF}\left(w_t|C_j\right)$ represents the total number of occurrences of the feature word w_t in all sample documents in the C_j class, and $\mathrm{TF}(w_t|C)$ represents the total number of occurrences of the feature word w_t in all categories.

Adding at as a weighting factor to the Bayesian formula, the classification function can be obtained as:

$$\begin{gathered} \arg\max \alpha_t P\left(C_j\right) \prod_{t=1}^{n} P\left(w_t|C_j\right) \\ C_i \in C \end{gathered} \tag{2}$$

where $P(C_j)$ is the probability of belonging to class C_j in the training text set, and $P\left(w_t|C_j\right)$ is the probability that the classifier expects the word w_t to appear in documents of class C_j.

4 Analysis and Research of Bayesian Classification Algorithm in Recognition of Insurance Tax Documents

4.1 *Implementation of Naive Bayes Algorithm*

After the data preparation stage before the experiment, the next step is to construct the Naive Bayesian classification model and the verification of the samples to be tested. The operation process is shown in Fig. 1.

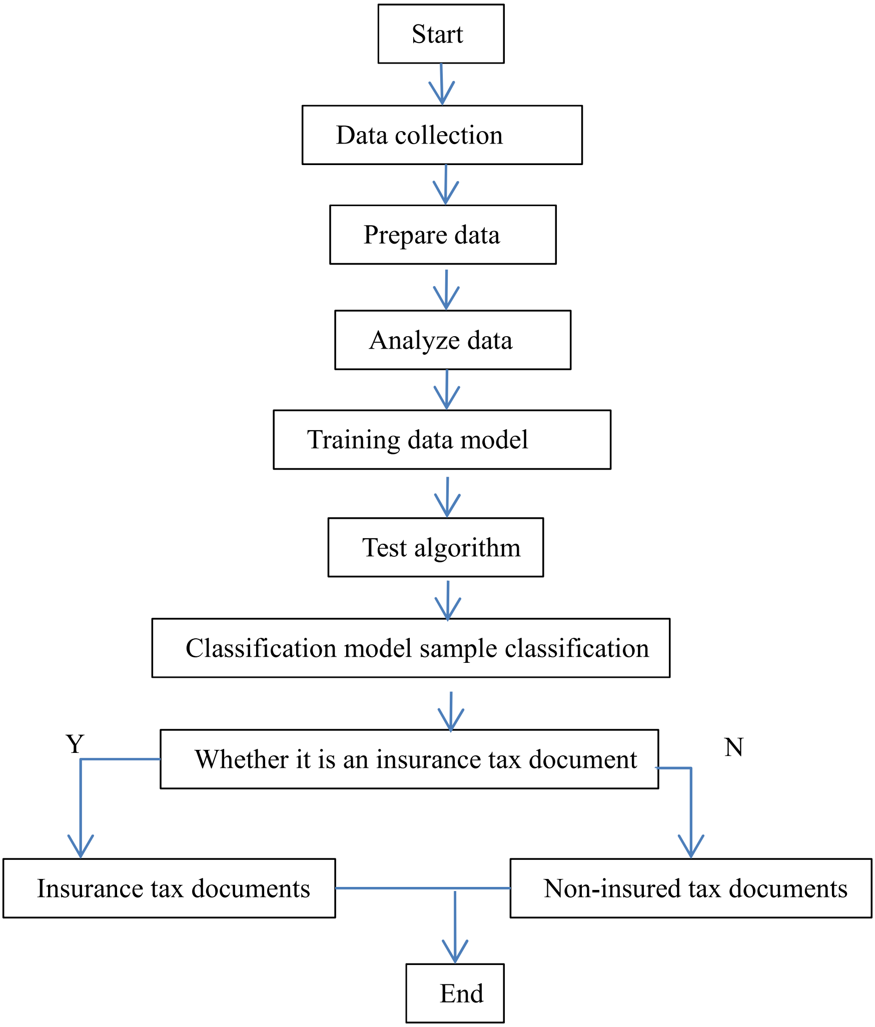
Start
Data collection
Prepare data
Analyze data
Training data model
Test algorithm
Classification model sample classification
Whether it is an insurance tax document
Y
N
Insurance tax documents
Non-insured tax documents
End

Fig. 1 Naive Bayes model construction process

Table 1 Comparison of experimental results

Experimental algorithm	Accuracy	Recall	$F1$ value
Naive Bayes	0.89	0.92	0.84
Weighted Naive Bayes	0.95	0.98	0.92

Based on the independent feature attributes of Naive Bayes, the experimental logic of the Naive Bayes classification model is clear. In the process of sample verification, first collect data, read the preprocessed sample dataset, then divide the data content into word vectors, train the classification model, integrate data features, and build a feature vector model database, and then, the obtained data model is used to detect the test data to realize the feature conversion of the data samples. Finally, the classification prediction evaluation index of the dataset is set, and the whole Naive Bayesian classification model is completed, and the experimental verification is carried out on the verified samples.

Based on multiple test experiments, the drawbacks of insufficient data samples are gradually revealed. Without a strong dataset as the support for building a training model, the data model cannot be fully trained, and many classifiers will appear in the sample classification stage. The database cannot identify the feature vector, so that the classifier has not been able to achieve high accuracy in the sample classification experiment and also greatly reduces the classification timeliness. This paper conducts an improved weighted Naive Bayes classifier experiment.

4.2 Analysis of the Improved Classifier Effect

The test results of the improved weighted Naive Bayes algorithm are given in Table 1.

According to the comparison data in Table 1 and Fig. 2, the improved Bayesian classifier is superior to the Naive Bayesian classifier in terms of accuracy, recall, and test value. The average accuracy of the improved weighted Bayesian classifier has reached 0.95, the recall rate has reached 0.98, and the $F1$ test value has reached 0.92, which are greatly improved compared to the original, as shown in Fig. 2. This is because some words appear frequently in texts of various categories, and it is difficult to judge which category they belong to by this word. The introduction of the at parameter reduces the degree of discrimination, thereby improving the degree of discrimination of low-frequency words, which better reflects the discriminative effect of vocabulary.

5 Conclusions

At present, my country’s insurance is still in its infancy, and there are very few tax policies specifically for insurance. It is particularly important to establish a special

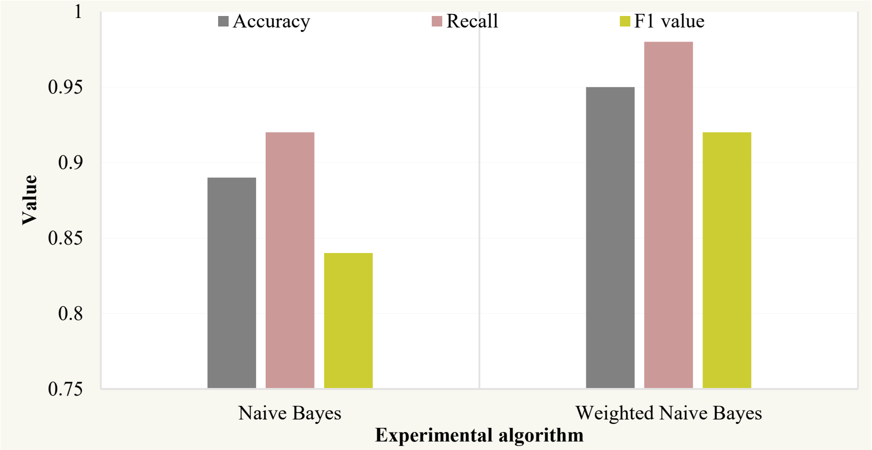

Fig. 2 Analysis of experimental results

document recognition system for the special text of official documents. In this paper, an experimental corpus of insurance tax documents is constructed, and combined with the characteristics of insurance tax documents, a Naive Bayesian classification algorithm and an improved Naive weighted Bayesian algorithm are used to design a text classifier for insurance tax documents, which has good results. However, the processing speed of the classifier still needs to be improved; the vocabulary of the dictionary set is not rich enough and needs to be continuously increased; the number of corpora is insufficient and needs to be continuously expanded. Due to the limited time, the effect of text length on classification was not involved in the experiment process, so the text length can be considered as an experimental factor; in order to achieve higher accuracy, the insurance tax official text library can be introduced; the semantic relationship between them is realized, and the second dimension reduction based on the ontology library is realized.

References

1. Ode S, Dowd B, Feldman R (2019) Comparing measures of physician market concentration using tax identification numbers versus independent negotiating units. Antitrust Bull 64(1):128–135
2. Aliev MA, Kunina IA, Kazbekov AV et al (2021) Algorithm for choosing the best frame in a video stream in the task of identity document recognition. Comput Opt 45(1):101–109
3. Susanto R, Putri FP, Wiratama YW (2018) Skew detection based on vertical projection in Latin character recognition of text document image. Int J Eng Technol 7(4):198–202
4. Ergn C, Norozpour S (2019) Farsi document image recognition system using word layout signature. Turk J Electr Eng Comput Sci 27(2):1477–1488
5. Al-Khaffat H, Musa NA (2018) Optical English font recognition in document images using eigenfaces. Innovaciencia 6(1):1–11