Research on Chinese Patent Text Classification Based on SVM

Ting Han[1]

[*] Corresponding author: h15530357553@163.com

[1]Shanghai Institute of Technology, School of Economic and Management, Shanghai 200030

Abstract In recent years, the substantial increase in the number of patent applications has brought great challenges to the classification of patent texts. In order to improve the efficiency of patent text classification and further improve the level of patent management. In this paper, a machine learning method based on SVM for Chinese patent text classification model is proposed. The research uses word vectors of jieba word segmentation and word2vec model for text representation, and uses five machine learning algorithms for text classification tests. After comparing the results, the SVM model is superior to other models in accuracy, recall, F1_score, etc. This research has important guiding significance for automatic classification of patent texts.

Keywords: Chinese patent, Text classification, word2vec, SVM, Machine learning

1. INTRODUCTION

In recent years, the number of patent applications in China has been on the rise. According to the statistics of the annual report of the State Intellectual Property Office of China, the number of patent applications in 2020 reached 5016,030, among which 1,344,817 were invention patent applications, 2,918,874 were utility models, and 752,339 were designs [1]. Such a large number of patent applications reflect the strengthening of our science and technology strength. In the face of the increasing number of patent papers, the manual classification of patent examiners alone cannot meet the needs of efficient and accurate classification of patent documents [2]. In order to further improve the efficiency of patent classification, reduce the workload of patent examiners and improve the accuracy of patent classification, the introduction of intelligent technology is of great significance for the automatic classification of patent text. Automatic classification of patent texts is of great significance for the issuance of new patents and patent retrieval. The precise classification of patent texts not only avoids a lot of manual repeated work, but also enables the applicant to avoid repeated research and patent infringement in a timely manner, thus generating huge economic value [3].

Automatic classification of patent text is a process in which the computer automatically assigns one or more patent classification numbers to patents according to specific rules, metadata, text content and other characteristics [4]. A large number of scholars at home and abroad have studied the algorithm of patent text classification. RICHTER G [5] used KNN method to automatically classify patent text, improving the classification accuracy from 70.8% to 75.4%; Hu Jie [6] and others proposed a patent text classification model based on convolutional neural network and random forest, which is superior to other single models in

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334244

English machinery patent classification. Li Sheng zhen [7] proposed an automatic patent classification method based on the back propagation neural network. In this study, IPC classification number was used as the patent classification standard, and the patent text with the patent classification as H02 was used as the data for testing, and good experimental results were obtained. The BERT-CNN multi-level patent classification model based on the pre training language model proposed by Lu Xiao lei [8] has an accuracy rate of 84.3%. Xiao Yue jun [9] and others proposed a Chinese patent text classification method based on feature fusion, which fused the BERT pre trained sentence vector with proper noun vector. The uneven distribution of data and the existence of a large number of unregistered words in the patent text have been improved. The above algorithms are applicable to different scenarios. For example, algorithms based on neural networks can adapt to large training sets, and KNN algorithm and BERT pre-training model can better deal with the problem of small sample classification. There are some combinatorial algorithms that are difficult to adjust parameters, and the effects will restrict each other.

To sum up, domestic and foreign scholars have conducted in-depth research on patent text classification from different perspectives and scenarios. Corresponding results have also been achieved. However, most of the current research on patent text classification focuses on English patents, and the research on the classification of a large number of Chinese patent texts is not yet mature. In order to improve the classification efficiency of Chinese patent texts and find the optimal classification model, this paper selects Chinese patent texts in the food field, and uses five machine learning classification algorithms to classify texts by building large-scale patent datasets, comparative analysis to select the best model. Finally, a new machine learning method is proposed, providing a reference for efficient and accurate patent classification of Chinese text.

2. RESEARCH METHOD

2.1 Overall framework

This paper is mainly to propose a better performance model to solve the problem of patent text automatic classification. Therefore, the research framework of this paper is shown in Figure 1, which mainly includes four steps.

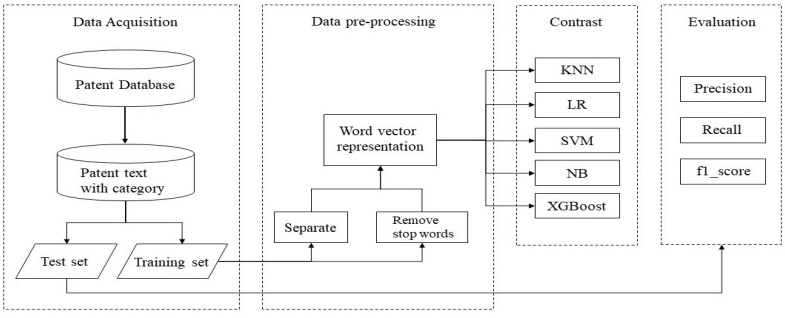

Figure 1. Framework of the classification method

2.2 Data acquisition

This paper selects the patent documents in the food field from the Chinese patent database of PatSnap as the corpus, and marks the data according to the International Patent Classification (IPC). Patent data in the corpus mainly includes classification number, title, abstract, sovereignty item, etc. According to the SooPAT IPC classification search results on the SooPAT website, Examples of the main classification symbols of patents in the field of food and the corresponding subject parts are shown in table 1. There are 10000 patent documents for A21 and A22 respectively, 5000 for A23B, A23D, A23F, A23J and A23K respectively, 8000 for A23C and A23G respectively, a total of 61000 patent documents. In the experiment, the content of the patent abstract is used as the corpus of the final experiment, and the data is divided manually into training sets and test sets according to 8:2 of each category. There are 48,800 pieces of data in the training set and 12,200 pieces of data in the test set.

Table 1. Patent classification number and corresponding subject matter

IPC main classification number	theme
A21	Baking; Equipment for making or handling dough; Baking dough
A22	Slaughter; Meat processing; Processing of poultry or fish
A23B	Preservation, such as storage of meat, fish, eggs, fruits, vegetables and edible seeds in cans…

2.3 Data preprocessing

The first step is to obtain data for manual annotation, followed by the preprocessing stage of patent text data. First, use jieba word segmentation to combine mechanical word segmentation with statistical word segmentation, divide the text content into several independent lexical units, and then delete stop words. In this study, "Harbin University of Technology Discontinued Words List" is used as the basis for deleting stop words, this study removed the words that affect the final classification effect. Using word vector embedding method and word2vec model, the semantic features of the text are extracted from the above data. The model will be constantly trained and optimized, the ultimate goal is to change the word into a vector format. The algorithm is trained in continuous bag of words (CBOW) and Skip gram model. In the actual operation process of this study, the Skip-gram model is more accurate than the CBOW model. In order to more accurately classify the effect, this paper selects the Skip gram model for training.

2.4 Machine learning method

In order to find the optimal classification model, this study used five machine learning classification algorithms for comparative analysis, including logistic regression, k-Nearest Neighbor, Naive Bayes model, Support Vector Machine and eXtreme Gradient Boosting. SVM algorithm is a widely used algorithm in machine learning. The combination of dimension reduction and classification has good generalization performance for classification problems [10]. The word2vec model is used to extract the semantic features of the text from the data, and the word vector is input into the SVM algorithm model for training to get a classification model. When the system needs to classify the new patent text, the patent text can be input into the model to obtain the classification results.

2.5 Evaluation indicators

The final model of the classification algorithm is the model with the minimum average error selected through cross validation. In this experiment, precision, recall and f1_score are used as evaluation indicators.

The accuracy rate is defined as the proportion of the predicted results as positive examples in the true positive samples:

$$Precision=\frac{TP}{TP+FP} \tag{1}$$

Recall rate is defined as the proportion of samples with positive prediction results:

$$Recall=\frac{TP}{TP+FN} \tag{2}$$

F1_score is an indicator of comprehensive accuracy and recall:

$$f1\_score=\frac{2\times Precision\times \mathrm{Re}call}{Precision+\mathrm{Re}call} \tag{3}$$

In the above formula, TP is the number of correctly classified samples, FP is the number of incorrectly classified samples, and FN belongs to this category but is wrongly classified.

3. EXPERIMENTS AND RESULTS

This experiment represents the text based on the word vector trained by word2vec, With the support vector machine model, the parameters of the model are constantly optimized and adjusted, and the Chinese patent classification model is obtained based on SVM. This experiment compares SVM with four classical classification models, LR, KNN, NB and XGBoost, and analyzes the accuracy, recall and f1_score of each model. The experimental results are shown in Table 2.

Table 2. Patent text classification results

classification	LR			KNN			NB			SVM			XGBoost		
	P	r	F1	p	r	F1	p	r	F1	p	r	F1	p	r	F1
A21	0.62	0.69	0.65	0.50	0.65	0.56	0.30	0.12	0.17	0.64	0.76	0.69	0.61	0.72	066
A22	0.63	0.89	0.74	0.53	0.83	0.64	0.31	0.95	0.47	0.63	0.91	0.75	0.62	0.89	0.73
A23B	0.73	0.64	0.68	0.63	0.52	0.57	0.47	0.16	0.23	0.74	0.68	0.71	0.73	0.64	0.68
A23C	0.76	0.54	0.63	0.70	0.47	0.56	0.55	0.18	0.27	0.82	0.56	0.67	0.78	0.55	0.65
A23D	0.93	0.83	0.87	0.87	0.71	0.78	0.50	0.75	0.60	0.95	0.84	0.89	0.95	0.80	0.87
A23F	0.91	0.85	0.88	0.84	0.74	0.79	0.79	0.35	0.48	0.94	0.86	0.90	0.94	0.84	0.89
A23G	0.66	0.61	0.63	0.65	0.44	0.52	0.61	0.06	0.11	0.76	0.63	0.69	0.68	0.61	0.65
A23J	0.87	0.73	0.80	0.85	0.67	0.75	0.63	0.65	0.64	0.91	0.76	0.82	0.90	0.73	0.81

A23K	0.96	0.95	0.96	0.94	0.94	0.94	0.70	0.92	0.79	0.98	0.95	0.97	0.97	0.95	0.96

From the experimental results, it can be seen that different classifiers have certain differences in the accuracy of the recognition of patent text categories. The accuracy of the five models from high to low is SVM, XGBoost, LR, KNN, NB. Among them, the NB has the worst accuracy for the classification of patent text, and the F1_score of each category is lower than the other four models. The F1_score of SVM model is higher than that of other models in each patent category classification, and the F1_score of A23K category is as high as 97%. It can be seen from the comparison in the table that the classification accuracy of the five classifiers for the A23K category is higher than that of other categories. The reason is that A23K has obvious grammatical topic characteristics and the data set is relatively complete. In the A23C category, the F1_scores of the five classifiers are low, indicating that the classification effect of this study on this category is average, and the SVM model is as low as 67%. The reason is that the syntax features of this type of patent text are somewhat similar to other patent texts. Through comparative analysis, in general, SVM based patent text classification is the best.

4. CONCLUSION

In the face of massive patent texts, a text representation model based on SVM algorithm using word vector of word2vec model is proposed. On this basis, the classification effect of SVM model is better than other models, and its accuracy, recall and F1_score are high. Although this research process mainly uses Chinese patent texts in the food field, the model mentioned in this paper can also be extended to text classification of patent categories in other fields. This paper first selects 61000 patent documents in the food field, constructs a large-scale patent data set, and uses the jieba word segmentation and word2vec model to represent the content of each text data. Through comparative analysis of five machine learning classification models, a Chinese patent text classification model based on SVM algorithm is obtained. It has guiding significance for improving the efficiency of Chinese patent text classification in the food field and achieving better patent text classification effect. However, there are still some shortcomings in this study. For A23C, a patent text with syntax characteristics similar to other patents, it is impossible to obtain better classification results. Next, we will use a better dataset to further study this problem.

Reference

[1] 2020 Annual Report on Intellectual Property Statistics. Patent Applications, Patent Grants and Patents In Force of Three Kinds Originated from Home and Abroad (2020) [R]. CNIPA, 2020.

[2] Bao Hai-long, Li Jin-lin. A Study of the Identify Method about IPC and Theme Terms on Patent Retrieval [J]. Journal of Beijing Institute of Technology(Social Sciences Edition),2003(05):74-76.

[3] Li S, J Hu, Cui Y, et al. DeepPatent: patent classification with convolutional neural networks[and word embedding[J]. entometrics, 2018, 117.

[4] Lyu Lu-cheng, Han Tao, Zhou Jian, Zhao Ya-juan. Research on the Method of Chinese Patent Automatic Classification Based on Deep Learning [J]. Library and Information Service,2020,64(10):75-85.

[5] RICHTER G,MACFARLANE A. The impact of metadata on the accuracy of automated patent classification [J] . World Patent Information,2005,27 (1) :13-26.

[6] Hu Jie, Li Shao-bo, Yu Li-ya, et al. A patent classification model based on convolutional neural networks and rand forest ［J］. Science Technology and Engineering,2018,18(6):268-272.

[7] Li Sheng-zhen, Wang Jian-xin, Qi Jia- dong, Zhu Li-jun Automated categorization of patent based on back-propagation network [J]. Computer Engineering and Design,2010,31(23):5075-5078.

[8] Lu Xiao-lei, NI Bin. BERT-CNN: A Hierarchical Patent Classifier Based on Pre-trained Language Model[J]. Journal of Chinese Information Processing,2021,35(11):70-79.

[9] Xiao Yue-jun, Li Hong-lian, Zhang Le , Lu Xue-qiang, You Xin-dong. Research on Chinese Patent Text Classification Method Based on Feature Fusion [J]. Data Analysis and Knowledge Discovery,2022,6(04):49-59.

[10] Dian Puspita Hapsari, Imam Utoyo, Santi Wulan Purnami, et al. Text Categorization with Fractional Gradient Descent Support Vector Machine[J]. Journal of Physics: Conference Series, 2020, 1477(2).