Research on Text Recognition System of Logistics Enterprise Policy Based on Text Mining

Jiahui Wang[1,a], Yidi Wang[1,b], Limei Xu[1,c*]

wangjiahui96628@163.com[a], 1119076056@qq.com[c*]

College of E-commerce and Logistics, Beijing Technology and Business University, Beijing, China[1]

Abstract: Modern logistics is an important support to realize the reform of supply side structure and the high quality development of economy. The Chinese government has also formulated a large number of policies to ensure the good and orderly development of logistics. Logistics enterprises concerned about the government policy can grasp the industry wind direction, good business decisions. However, logistics industry is a complex industry, with a wide range of policies, a large number of policies and complex contents. Therefore, enterprises are prone to omissions or inadequate grasp of key fields in practice. Therefore, this paper designs a policy text recognition system for logistics enterprises based on text mining. TF-IDF algorithm is used to extract the feature words of policy texts, and random forest is used to classify policy texts. The results are compared with manual labeling results to calculate the accuracy and recall rate. Through experiments, it is found that random forest algorithm has a high accuracy rate of policy text recognition. Logistics enterprises can use random forest algorithm to identify and analyze policy text, so as to improve the working efficiency and decision-making accuracy of logistics enterprises.

Keywords: Text Mining, Logistics Policy, Analysis System.

1 INTRODUCTION

The total amount of social logistics increased from 219.2 trillion yuan in 2015 to 335.2 trillion yuan in 2021. Logistics has become a strong growth point of the national economy and an important driving force to promote economic development. Especially in the current situation, it is of great strategic significance to vigorously support logistics development to cope with the challenges of economic development under the epidemic, and to establish a new development pattern with the domestic cycle as the main body and the domestic and international cycles mutually reinforcing. With the rapid development of Internet technology and the growing maturity of new generation of basic technologies such as big data and the Internet of Things, logistics has also merged with other industries to give birth to various new models. However, because of the complex property of the logistics industry, the development of the logistics industry was in the state of disorder for a time, which aroused the government's strong

BDEDM 2023, January 06-08, Changsha, People's Republic of China

DOI 10.4108/eai.6-1-2023.2330301

attention. In 2009, The State Council pointed out in the Logistics Industry Adjustment and Revitalization Plan that the legislative research and policy formulation of the logistics industry need to be strengthened, and the corresponding policy and regulation system needs to be improved and improved.

After more than ten years of development, government departments at all levels continue to introduce various policies to guide and support the logistics industry. Studies have proved that the time required for logistics enterprises to adapt to the policies is getting shorter and shorter. Meanwhile, the implementation of the policies has a positive effect on the stock market, market value of logistics enterprises and the development of the logistics industry [1-2]. However, the current policies are numerous, complex, and involve a wide range of industries. Most departments of logistics enterprises collect and apply the corresponding policies manually, which is inefficient and easy to be neglected or repeated between departments in practice. They cannot keep up with the national policy direction and miss some subsidies and support measures. To some extent, this will also affect the decision-making accuracy and future development of logistics enterprises. Therefore, in view of the characteristics of the current logistics policy update fast, large quantity and complex industry, it is necessary to explore the automatic capture and distribution of policy text, establish real-time identification and coordination mechanism, which is conducive to better decision-making and practice of logistics enterprises.

In the era of big data, text data is also growing exponentially. How to capture valuable information in numerous text data is also the research direction of many scholars. Feldman put forward the concept of text mining in 1995 [3]. Text mining mainly uses machine learning techniques such as decision tree and deep neural network to extract text information and analyze text by training machines [4]. After a lot of practice, text mining is widely used in social media because of its high accuracy and speed in text feature analysis, cluster analysis, text emotion and topic extraction [5], policy analysis [6] Literature metrology [7] and other fields of study. Tang Heng [8] textrank method was adopted to extract keywords in the text of intellectual property policies for smes and calculate the weight of each keyword, which directly reflects the focus of intellectual property policies for smes. Ding Siyuan [9] According to the content characteristics and discourse characteristics of public hearing texts, et al. put forward a three-stage event extraction method to realize the extraction of valuable information and conduct in-depth analysis of the extracted valuable information.

In summary, guided by the needs of logistics enterprises for policy interpretation, this study intends to propose an identification mechanism of text classification and feature extraction based on text mining to conduct feature classification and automatic analysis of policy texts, so as to improve the attention and correctness of logistics enterprises to policies, and help logistics enterprises to make correct decisions and develop business strategies consistent with the national wind direction.

2 DESIGN OF POLICY TEXT RECOGNITION SYSTEM

The policy text identification system of logistics enterprises is responsible for automatic identification and distribution of policy texts, and value analysis and processing after processing by various departments, as shown in Figure 1.

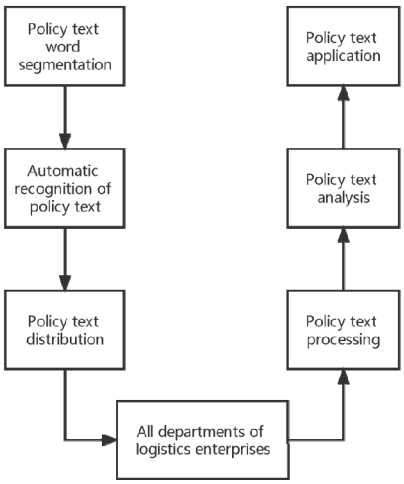

Figure1. Policy text recognition system of logistics enterprises

The policy text recognition system of logistics enterprises is based on text classification, including data crawling and text feature analysis. Due to the complex nature of logistics industry, policy texts come from many sources and involve a wide range of contents. If manual collection and analysis are carried out, the efficiency is low and the needs of different departments are different, resulting in poor communication. Therefore, this system is based on python and adopts crawler program to collect policy text data, which can establish a general logistics policy text database for logistics enterprises. The policy texts of logistics enterprises are mainly from the state and government functional departments at all levels, such as provinces, cities and counties.

```
print('Grabing data....')
datas = list()
for u in urls:
    try:
        res = requests.get(u,headers=headers)
        html = res.text
        soup = BeautifulSoup(html,'html.parser')
        data = {}
        data['url'] = u
        data['title'] = soup.h1.string
        data['time'] = soup.find('time',itemprop=
        if soup.find('div',id='article_deck') is
```

Figure2.Code used for data crawling (part)

The policy text identification system of logistics enterprises mainly has the following steps:

Firstly, the original policy text data is preprocessed for word segmentation before extraction and analysis. In this paper, the current mainstream Chinese word segmentation tool - jieba segmentation is adopted for word segmentation. In order to ensure the accuracy and comprehensiveness of word segmentation, this paper uses "HIT University of Technology Glossary of Words" to remove words of words, and introduces "Standard Terms of Logistics (GBT18354-2021)" to build a dictionary of special terms for logistics. In feature engineering, the word bag model is used to represent the text in vector form. The word bag model constructs all the entries in the text data set into a dictionary, and represents each text as a frequency set of entries. The crawler program was used to collect the relevant policy text data, delete the invalid information and use the constructed dictionary for word segmentation of the original policy text.

The TF-IDF algorithm is used for keyword analysis of the policy text data after cleaning, and the category differentiation ability of the entry is explained by calculating the word frequency and reverse file frequency, so as to determine whether the entry is the keyword of the text. It tends to filter out common words and retain important words.

Where, TF is word frequency, which represents the frequency of entry appearing in the document; IDF is the reverse file frequency, indicating the frequency of the file in which the term appears in the whole file set. Multiply the two values together to get the TF-IDF value of the word. The greater the TF-IDF value of the word, the more important it is to the article. It can be expressed by the following formula:

$$TF = count(tj) / count(dj) \quad (1)$$

$$IDF = logN / num(t) + 1 \quad (2)$$

$$TF\text{-}IDF = TF * IDF \quad (3)$$

Where, count (tj) represents the number of words t contained in document j; count (dj) represents the total number of words contained in document j; N is the total number of documents; num (t) represents the number of documents containing words t.

The TF-IDF value of each word in the processed database is calculated one by one, and the TF-IDF value set form of each word is used to represent each document, and the irrelevant data is removed, and finally a matrix form document set is obtained.

After understanding the relevant policy texts, this paper takes 106 logism-related policy texts issued by the official websites of The State Council and its subordinate departments from January 1, 2018 to December 31, 2021 as data sources, and marks the department keywords according to the characteristics of logistics enterprises, as shown in Table 1.

Table1 Department - Keywords in the text of logistics policy

The serial number	department	Labeling Keywords	Number of texts
1	Administr-ation and Personnel Department	Talent recruitment, vehicle management	14
2	Finance Department	Tax administration, corporate income tax, value-added tax, subsidies	52
3	Management Department	Enterprise strategy, logistics park, pilot, key planning, cultivation, industrial chain, operation mode	98
4	Department of Transportation	Combined transportation, airport, hub layout	76
5	Warehouse management Department	Inventory, material turnover, agglomeration, cross-border	69
6	Research and Developme-nt Department	Technical support, network platform, standardization, infrastructure, intelligent logistics equipment	43
7	Marketing Department	E-commerce, rural logistics, application	87

Classification model can map data records in the database to a given category, including decision tree, logistic regression, naive Bayes, neural network and other algorithms. As a newly emerging and highly flexible machine learning algorithm, random forest has the highest accuracy among all the current algorithms. Besides, it is able to evaluate the importance of various features in classification problems and process input samples with high-dimensional features, so it has obvious advantages in estimation and inference mapping [10]. CART classification regression tree is a typical binary decision tree. In this paper, CART tree is used as a weak classifier, and sklearn standardization tool is used to establish and train the classifier.

3 EXPERIMENT OF POLICY TEXT IDENTIFICATION OF LOGISTICS ENTERPRISES

80% of the 106 logistics policy texts collected were used as training texts to establish and train the classifier, and 20% were used as test texts to test the effect of the classifier. The main steps include: segmentation of training data and test data, extraction of dictionary feature vector-quantization, and random forest prediction.

```
import collections
import pickle
from tools import load_data,word_to
from sklearn.model_selection import
from sklearn.feature_extraction imp
from sklearn.ensemble import Randon

all_data = load_data("train.xlsx")

all_content = list()
for content in all_data["content"]:
    temp = dict()
```

Figure 3 Model training code (part)

In the random forest model, mtry and ntree are two important parameters.mtry is the number of variables contained in each decision tree, which is generally tried one by one to achieve a better value.ntree is the number of base classifiers included. Generally, the value when the error in the model is stable can be roughly judged by the graph. Through the experiment, it is found that a better result can be obtained when the mtry value is 7.The setting of ntree should make the overall error rate of random forest stable, so the value of ntree should be large enough to ensure RF convergence. The results show that the classification effect is best when ntree=20.The experimental results are shown in Table 2.The text classification model based on random forest algorithm achieves the best classification effect when mtry=7 and ntree=20.

Manual labeling has been carried out in the previous paper, and the comparison with the results of the test text classification by the classifier shows that the accuracy rate of random forest classification is above 98%. In order to further verify the accuracy of the classifier, this paper also uses the departmental recall rate for evaluation. The results are shown in Table 3.

Departmental recall rate

$$R_i=x_iy_i/\sum_{i=1}^{7}x_iy_i \qquad (4)$$

The value of i ranges from 1 to 7 for each of the seven departments.

Table 2 Experimental results

ntree	precision
10	0.744
15	0.753
20	0.761
25	0.757

Table 3 Classifier accuracy rate and department recall rate

The serial number	Departm-ent	Correct quantity	Accuracyrate	Rate of recall
1	Administration and Personn-el Departm-ent			0.982252
2	Finance Departm-ent			1.000000
3	Managem-ent Departm-ent			0.970874
4	Departm-ent of Transpo-rtation			0.983951
5	Warehou-se managem-ent Departm-ent			0.964497
6	Researc-h and Develop-ment Departm-ent			0.983696
7	Marketi-ng Departm-ent			0.995634
A combi-ned		102	0.980769	

It can be seen from Table 2 that the accuracy of random forest is relatively high, and the classification accuracy of financial department policy text even reaches 100%. It can be considered that random forest can be practiced in the policy text recognition system of logistics enterprises.

4 CONCLUSION

Based on the complex characteristics of logistics industry and from the perspective of logistics enterprises, this paper uses text mining technology to put forward the design idea of policy text recognition system of logistics enterprises, and uses random forest to conduct classification experiment. The experimental results show that this algorithm is effective for the recognition of policy text and department matching of logistics enterprises. The system has a certain practical value, logistics enterprises timely grasp the relevant policies, adjust the company's operation direction, make the right strategic decision plays an auxiliary role, help logistics enterprises to develop in high quality. In the future research, the keywords can be subdivided and improved according to the needs of various departments of logistics enterprises, and more data sets can be introduced for training. Compared with other algorithms, the identification accuracy rate and department matching effect can be better.

REFERENCES

[1] Zhu Xiangping. The effect of smart logistics policy implementation on shareholder value: Based on the empirical data of Shanghai and Shenzhen A-share listed logistics enterprises [J]. Business Economics Research,2022(11):101-104.]

[2] Zou X. Empirical study on the impact of "Replacing business tax with value-added Tax" on the logistics tax burden -- based on the empirical data of Shanghai and Shenzhen A-share listed companies [J]. Journal of Zhongnan University of Economics and Law,2016(02):43-47.

[3] Feldman,R.Dagan.KDT-Knowledge Discovery in Textual Databases[A].In:Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining,1995:112-117.

[4] Wang Jianxin, Wang Ziya, Tian Xuan. An overview of natural scene text detection and recognition based on deep learning[J]. Journal of Software,2020,31(05):1465-1496.DOI:10.13328/j.cnki.jos.005988.

[5] Shi Shan-chong, ZHU Ying-nan, ZHAO Zhi-gang, KANG Kaili, Xiong Xiong. Investor Sentiment and Stock market Performance based on wechat text Mining [J]. Systems Engineering Theory & Practice,2018,38(06):1404-1412. (in Chinese)

[6] Qi Chang. Industrial Policy Analysis Based on Text Mining [D]. Dongbei University of Finance and Economics,2022.

[7] Liang Shuang, Liu Xiaoping. Research Progress on topic evolution of scientific literature based on Text mining [J]. Library and Information Services, 2002,66(13):138-149.

[8] Tang Heng, HD, Sun Yinglin, Xiao Hanzi. Research on smes' Intellectual property Policy Based on Text Mining -- Data from the central Level [J]. Science and Technology Management Research, 2002,42(01):92-100.

[9] Ding Siyuan, Qiao Xiaodong, Zhang Yunliang. Research on Public Hearing text Mining Method based on Event Extraction technology [J]. Journal of Information,202,41(01):52-59+30.

[10] Luo X. Research on text classification model based on Random forest [J]. Journal of Library and Information Science for Agriculture,2016,28(11):50-54.