International Journal of Innovation Scientific Research and Review
Vol. 03, Issue, 11, pp.1958-1960, November, 2021
Available online at http://www.journalijisr.com
SJIF Impact Factor 4.95

IJISRR

ISSN: 2582-6131

Research Article

AUXILIARY EVALUATION SYSTEM DESIGN FOR TARGETED POVERTY ALLEVIATION

ChaoYang, * ZhanWen, YanheNa, XuanLiu

School of Communication Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, China.

Received 14th September 2021; Accepted 15th October 2021; Published online 20th November 2021

ABSTRACT

Targeted poverty alleviation refers to a poverty control method that uses scientific and effective procedures to accurately identify, assist and manage the poverty alleviation objects according to the environment of different poverty-stricken areas and the situation of different poor farmers. However, in the implementation of targeted poverty alleviation policy, whether the applicant is a poor household mainly depends on the subjective judgment of grass-roots poverty alleviation staff and the historical data analysis model of previous years. In such a situation, it is inevitable that there are some phenomena of pretending to be poor households. Because this method relies too much on people's subjective judgment, a personal economic level prediction model based on data analysis and machine learning technology to help poverty alleviation workers identify poor households came into being. The model can effectively solve the defects of traditional methods.

Keywords: targeted poverty alleviation, machine learning, text data analysis, crawler.

INTRODUCTION

The research status of economic level prediction model is generally divided into three categories:

1. According to the forecast scope, there are national economic forecast, enterprise economic forecast, departmental economic forecast and regional economic forecast between the two, as well as world economic forecast.
2. According to the timeliness of prediction, there are short-term prediction, long-term prediction and medium-term prediction.
3. According to the nature of prediction, there are qualitative prediction and quantitative prediction.

However, these three methods are mainly applied to the economic level prediction of regions or groups, and there is still a lot of research space for the individual based economic level prediction model. This study takes user with different economic levels as the experimental object, collects the text data of user with different economic levels, and establishes a machine learning model based on these data. In the research process, firstly, the collected original data are analyzed and processed to facilitate the training of machine learning model. Secondly, the data are processed through the machine learning model, and then the prediction results are compared with the actual results to select the best model parameters that can accurately predict the personal economic level. The text data is collected from Sina Weibo (an open social platform in China). The main purpose of this study is to help the poverty alleviation staff preliminarily identify the applicant's personal economic level, so as to facilitate the follow-up poverty alleviation work. The rest of this paper is organized as follows: the second part introduces the similarity research, and introduces the methods and theoretical knowledge in detail. The third part introduces the experimental process. The fourth part analyzes and summarizes the experimental results.

***Corresponding Author: ZhanWen,**
School of Communication Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, China.

ECONOMIC FORECASTING MODEL STRUCTURE

Data preprocessing

Chinese word segmentation

Word segmentation is the process of recombining continuous word sequences into word sequences according to certain norms. As we know, in English writing, spaces are used as natural separators between words, while Chinese is only words, sentences and paragraphs, which can be simply delimited by obvious separators, except that words do not have a formal separator. Although English also has the problem of phrase division, at the word level, Chinese is much more complex and difficult than English.

Keyword extraction

The keyword extraction API provides an interface for extracting keywords. Through this API, the core content that the text wants to express can be extracted from a large amount of information. It can be entities with specific meaning, such as person name, place, film, etc., or some basic but key words in the text. Through this API, the extracted keywords can be sorted from high to low according to the weight in the text. The higher the ranking, the higher the weight, and the more accurate the core content of the text can be extracted.

Remove stop words

In information retrieval, in order to save storage space and improve search efficiency, some words or words will be automatically filtered out before or after processing natural language data (or text), which are called stop words. These stop words are manually input and not automatically generated. The generated stop words will form a stop word list.

Text Vectorization

Computers can't understand human language, but computers can understand numbers. Text vectorization is to represent text as a series of vectors that can express text semantics. It is an important way of text representation. Text representation is the basic work in natural language processing. The quality of text representation

directly affects the performance of the whole natural language processing system.

Principal component analysis (PCA)

Principal component analysis (PCA) is a feature dimension reduction method. It replaces the original features with a series of concise new features [3]. These new features are linear combinations of the original features. It maximizes the sample variance and tries to make the new features irrelevant. In this way, it is convenient to study and analyze the influence of various features on the model, effectively reduce the complexity of the model and improve the speed of training the model.

Machine learning model

As for the machine learning model, we choose the logistic regression model. It has many advantages. When classifying, the amount of calculation is very small, the speed is very fast, and the storage resources are low.

EXPERIMENTAL PROCESS AND RESULT

Experimental process

Based on the general process of machine learning training data, the experimental process of this paper is as follows.

- Collect the text data of people with different economic levels through the crawler program.
- Classify and store the collected text data and start the word segmentation step.
- Using the 11 categories of the open Chinese Thesaurus of Tsinghua University, the keywords of each article are divided into 11 categories through the word bag model, and the excel table is generated
- The processed vector (or PCA vector) without dimensionality reduction is used to train the logistic regression model.

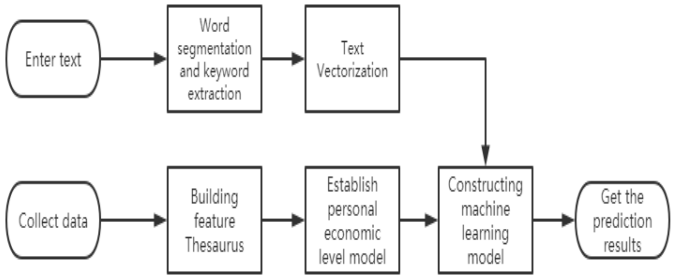

Fig 1: Processing process

Data acquisition

The data used in this experiment is to crawl the text data of users with different economic levels of microblog through Python crawler program. When crawling the data, we should pay attention to distinguish the text information of ordinary people and rich people.

Data preprocessing

After crawling the text, classify it and store it first. Because the original data is text, it is necessary to segment words into individual words, and remove the stop words in the text. It is still similar and meaningless words. Word segmentation uses Python Jieba word segmentation package.

Word segmentation

After that, the topn keywords are extracted from each document (n can be customized in the program according to the actual situation). These keywords are selected according to the TF IDF value and can be called through the python package Jieba. Keywords can represent each document, that is, the words that can best represent the person's position, opinion and emotion in the words and articles published by each microblog user.

Text Vectorization

After extracting the keywords, the txt file (keyword txt) representing everyone is transformed into digital form by using the word bag model. Tsinghua University thesaurus divides vocabulary into many types. Write a program to compare each keyword TXT document with Tsinghua University thesaurus to see which category the keyword data belongs to and count under which category of vocabulary (word frequency statistics). Each person's emphasis on speaking and publishing text is different, and the keywords are different. By counting the frequency of keywords in each type of document, we can get the document vectorization representation of each person (each document).

Principal component analysis

After obtaining the vectorized data (data in digital form), each row of the table is a document corresponding to a person. If it is a rich person, the label after the data is 1 and the label of ordinary economic population data is 0. A PCA (principal component analysis) is performed on the vector to compress the data and prevent over fitting.

Model training

Input the data into the model. This experiment adopts the logistic regression model, which is called by sklearn machine learning package (Python).

Result display

The final display of this experiment is to provide a user interface through the web page. Users can input the applicant's application text into the corresponding area to get the corresponding economic level prediction results.

Experimental result

After running the program, the program will segment the input text, and give the result as follows.

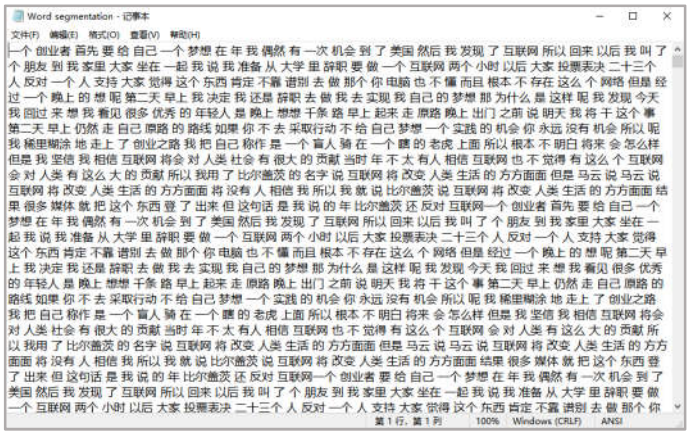

Fig 2: Word segmentation result

Then the program extracts the keywords from the word segmentation results, and give the result as follows.

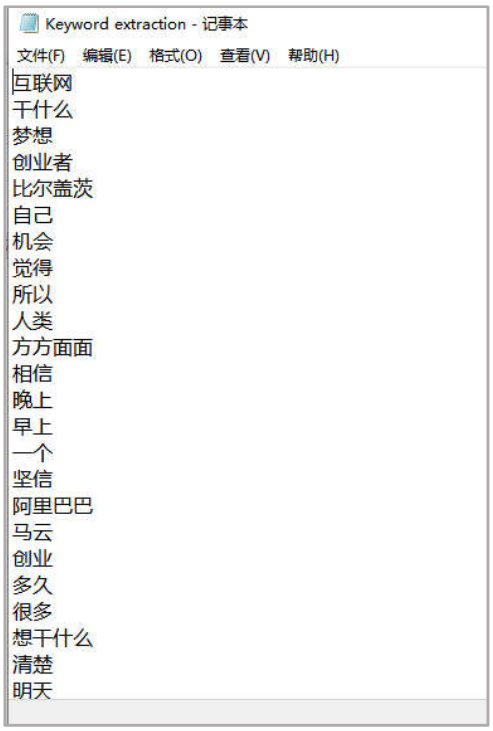

Fig 3: Keyword extraction results

After the second step of vectorization processing, the results are as follows.

```
F:\通信184-10组-程序源码-整个可运行程序目录\project1\106\gjc\guanjianci1.txt:Data vectorization complete

Process finished with exit code 0
```

animal	Economics	car	idion	place name	food	IT	law	Historical names	medicine	Poetry
0	10	0	0	2	0	2	0	2	1	0

Fig 4: Text vectorization results

Import the vectorized data into the machine learning model, and then obtain the prediction results and prediction accuracy.

```
Test set prediction results: [0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0
 1 0 0 0 0 1 1 1 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 1
 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0
 1 0 0 0 1 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 1 1 1 0 0 1 0 1 0 1 0 0 0 1 0]
Accuracy: 0.8055556
Press any key to continue
```

Fig 5: Test set prediction results

The data collected in this experiment are mainly the rich and ordinary people. The data collection for the poor is not very smooth, which is also one of the problems to be solved in the follow-up of this study.

CONCLUSION

The data collected in this experiment are mainly the rich and ordinary people. The data collection for the poor is not very smooth, which is also one of the problems to be solved in the follow-up of this study. The system is a precision poverty alleviation auxiliary evaluation system designed by combining psychology and machine learning algorithm. The system uses web crawlers to collect the online text data of people at all economic levels, carries out text processing through Chinese word segmentation, keyword extraction, text vectorization and other methods, constructs a feature thesaurus, establishes a machine learning model, and finally develops a mobile app to provide an input interface. When inputting the text data related to the poor population, the system can obtain the corresponding economic level analysis report of the poor household, assist the poverty alleviation staff to verify the identity of the poverty alleviation object, and better promote the targeted poverty alleviation work.

ACKNOWLEDGEMENTS

This work is supported by College Students' Innovative Entrepreneurial Training Project, Chengdu University of Information Technology（202010621203）.Thanks to the students who worked hard to collect data in this project and the teachers who guided patiently.

REFERENCES

1. Multilingual Phone Recognition in Indian Languages - Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning. Springer 2022, ISBN 978-3-030-80740-5, pp. 1-88
2. Wang Ding. Analysis and Research on natural language processing technology [J]. Introduction to scientific and technological innovation, 2020,17 (07): 141-142
3. Feng Shaodi. Natural language processing technology based on Reinforcement Learning [J]. Digital world, 2020 (03): 9-10
4. Chen Yunchuan, song Hao, Zhao Ye, Liu fawen. Data mining and application of Campus All-in-one Card Based on logistic regression algorithm [J]. Journal of Kunming metallurgical college, 2020,36 (03): 57-61
5. Python, Data Science and Machine Learning - From Scratch to Productivity. World Scientific 2022, ISBN 9789811215728, pp. 1-300
6. Classification and prediction of bulk densities of states and chemical attributes with machine learning techniques. Appl. Math. Comput. 412: 126587, 2022.
7. Modeling Natural-Image Spaces for Single-Label Image Classification & Photo-Realistic Style Transfer and Directionally Paired Principal Component Analysis for the Estimation of Coupled Data. Georgia Institute of Technology, Atlanta, GA, USA, 2021.
8. Fast Circulant Tensor Power Method for High-Order Principal Component Analysis. IEEE Access 9: 62478-62492, 2021.

****