Research on Text Communication Security Based on Deep Learning Model

Guanghua Yu[1,2(✉)] and Wanjuan Cong[1,2]

[1] Heihe University, Heihe, Heilongjiang, China
Ygh2862@163.com
[2] School of Computer and Information Engineering College, HeiHe University, Heihe, China

Abstract. In response to the current spam flooding problem, this paper uses Python language machine learning and natural language processing technology to study the identification classification of spam messages. The Jieba algorithm is used to distinguish the Chinese word, and the TF-IDF algorithm is used to conduct feature extraction. On the basis of the analysis of the classifier algorithm, the experimental data is finalized. The results show that the classification effect of the polynomial plain Bayes classifier is optimal, and the identification of garbage text is best optimized.

Keywords: Spam message · Naive Bayesian Model

1 Preface

With the wide application of social media, text message service has developed rapidly. At the same time, a large number of junk information and fraud information are added to it, which brings different degrees of harassment and serious security risks to people's lives, and adds unstable factors to the development of a harmonious society. Therefore, the use of information technology to establish the ability to identify, correct and deal with spam text information is particularly important. In this paper, the deep learning algorithm and python language are used to identify and classify spam text messages. Jieba Chinese word segmentation tool and TF-IDF feature extraction are used to analyze the principles of naive Bayes, Gaussian distribution, random forest and other classifier algorithms. Finally, according to the experimental data, the classification effect is compared [1].

2 Correlation Theory

There are three methods to effectively identify spam messages, which are black and white list method, rule-based method and SMS content-based method [2–5]. The blacklist and rules is relatively simple, but the disadvantage is that the number list and keywords need

Foundation item: School-level topics (KJZ202102); School-level topics (XJGY201923) Project of Heilongjiang Provincial Department of Education (2019-KYYWF-0462).

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 147–154, 2022.
https://doi.org/10.1007/978-3-031-04409-0_14

to be added manually, and the number that can be added is relatively limited and difficult to be comprehensive, resulting in poor recognition effect. In view of the limitations of the two methods, the current research on spam SMS identification technology mainly focuses on the content of SMS, using text classification technology to transform spam SMS recognition problem into a supervised learning problem. Text classification technology is based on machine learning algorithm. Firstly, it extracts the features of the manually marked text, and then using the algorithm to classify the text automatically.

2.1 Chinese Participle

In text data mining, word or phrase is usually used as feature to segment words. Therefore, we need to extract the original text data by word segmentation to obtain the corresponding feature list before text classification and feature extraction. Therefore, high accuracy of text segmentation has a great impact on the subsequent text analysis and text classification. At present, English and Chinese are the main text segmentation methods. Due to the different grammatical structures, the two languages have different word segmentation methods. However, English word segmentation is much easier than Chinese word segmentation in practical point, because English word results and punctuation are relatively clear, while Chinese text is relatively vague. This paper uses Chinese text, and uses Jieba Chinese word segmentation to segment Chinese text.

1 The construction principle of prefix dictionary: Prefix dictionary is constructed by using statistical dictionary algorithm. For example: “Heihe College” is a word in the statistical dictionary. The prefix of the word is {“Hei”, “Heihe”, “heihexue”}, and the prefix of the word “College” is {“Xue”}. By analogy, we can get the prefix Dictionary of the Chinese text to be segmented.
2. The construction principle of directed acyclic graph based on prefix dictionary is shown in Fig. 1.

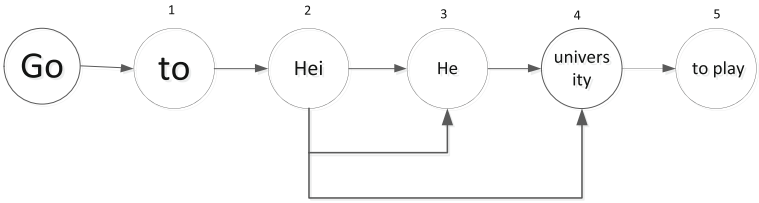

Fig. 1. An example of directed acyclic graph based on prefix dictionary

From Fig. 1 can be seen, there are three ways to divide “Hei” into “Hei”, “Heihe” and “Heihe College”, while for “Qu”, there is only one way to divide “Qu”. The reason is that the word “Qu” has no prefix. Similarly, the combination of “Xue” and “play” can be obtained. Based on this, a directed acyclic graph is constructed.

3. The algorithm principle of using dynamic programming algorithm to find the maximum probability path: the maximum probability path of the front drive node must be calculated for every node found by the dynamic programming. The core condition of dynamic programming algorithm is to have repeated subproblems and optimal substructures. For the problem of finding the maximum probability path in directed acyclic graph, the repetitive subproblem and the optimal substructure are as flows:

(1) Repetitive subproblem: In a directed acyclic graph, any node di may have some successor nodes dj or dk, and it is necessary to repeatedly calculate the probability value of the path to di for each successor node. The mathematical expressions are 1 and 2.

$$P(i) -> j \ = \ P(i) + \text{weight}(j) \tag{1}$$

$$P(i) -> k \ = \ P(i) + \text{weight}(k) \tag{2}$$

(2) The optimal substructure is as follows: For the end node dx of a directed acyclic graph, there may be multiple precursor nodes di and dj. The maximum probability paths for DK to reach these precursor nodes are respectively $P_{max(i)}$, $P_{max(j)}$ and $P_{max(k)}$. According to this, the maximum probability path is $P_{max\,(x)}$, and the calculation formula is 3.

$$P_{max(x)} = \max\left\{\left(P_{max(i)}, P_{max(j)}, P_{max(k)}.....\right) \ + \ \text{weight}(dx)\right\} \tag{3}$$

2.2 Feature Representation and Extraction of Text Data

TF IDF algorithm is used to express and extract text data features, that is, the frequency of a word appearing in text. Because word frequency has a great influence on the classification of original text data, the greater the word frequency of a word, the greater its contribution to text recognition. The frequency of words is shown as 4:

$$\text{tfij} = \frac{\text{nij}}{\sum \text{nkj}} \tag{4}$$

nij is the number of times that the word appears in the file dj, and the denominator is the total number of times that all words appear in the file dj.

$$\text{idf}_j = \log(|D|/|\{k{:}t \in dk\}|) \tag{5}$$

|D| is the total number of files in the corpus. |{k:t ∈ dk}| denotes the number of files containing the word ti (i.e. the number of files with ni, j ≠ 0).

The TF-IDF algorithm process takes five groups of Chinese text data as examples, as shown in Table 1:

1. Calculate the inverse document frequency of Chinese text data: The algorithm counts the number of different words appearing in the text. For example, from the text data

Table 1. Five groups of Chinese text data content

Number	Content
1	Pets have pets, pets, pets, pets, pets
2	Pets include dogs, cats, hamsters, hedgehogs and squirrels
3	The animals are lovely. I like lions best
4	Dogs are loyal pets
5	Lovely pets are dogs, cats, chinchillas, other pets are hamsters and lizards

content in Table 1, we can see that "pet" is quoted in text 1, 2, 4 and 5, it appears in 4 places, so the reverse document frequency of "pet" is 4. Similarly, the reverse document frequencies of other nouns are: pet = 4, dog = 3, cat = 2, hamster = 2, cute = 2, hedgehog = 1, beast = 1, lion = 1, loyalty = 1, chinchilla = 1, lizard = 1. Therefore, "pet" and "dog" are the two most important words in the five sets of texts. Remove text 1 and text 3.

2. Calculate the word frequency of Chinese text data: After the algorithm calculation of inverse document frequency, text 2, text 4 and text 5 are selected. Because the frequency of words in text has a significant impact on text classification, "pet" is the core word in the text. Through calculation, it is found that the word appears twice in the fifth text, and only once in the second and fourth text. Therefore, the final ranking result is shown in Table 2.

Table 2. Five groups of sorted Chinese text data content

Sort	Content
1	Lovely pets are dogs, cats, chinchillas, other pets are hamsters and lizards
2	Pets include dogs, cats, hamsters, hedgehogs and squirrels
3	Dogs are loyal pets
4	Pets have pets, pets, pets, pets, pets
5	The animals are lovely. I like lions best

2.3 Segmentation of Training Set and Test Set

The paper uses train_test_split method of sklearn.model_selection is used to segment the data. According to 75% and 25% segmentation ratio, the text data is divided into training data set and test data set.

3 Naive Bayesian Recognition Model

Firstly, the text data is extracted, and then the prior probability belonging to the feature value is calculated. According to the obtained prior probability, the Bayesian formula is

used to calculate the posterior probability. The flow chart of naive Bayesian algorithm is shown in Fig. 2.

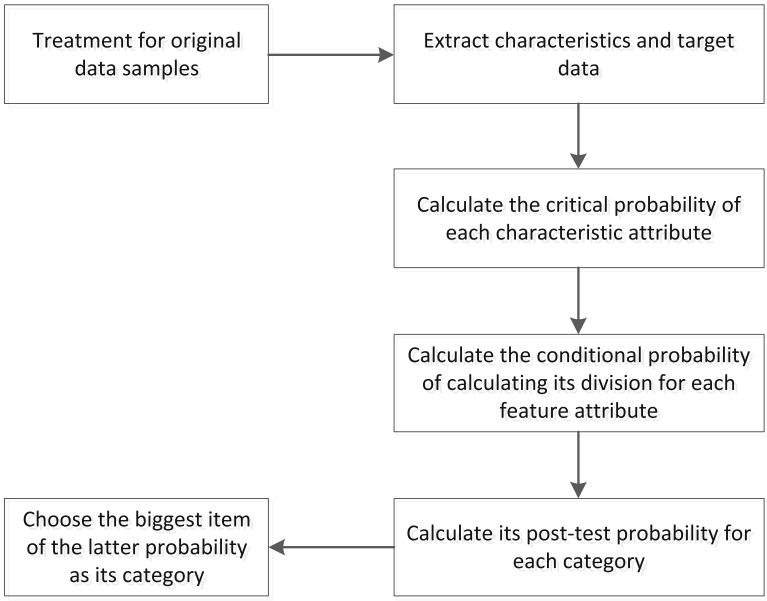

Fig. 2. Flow chart of naive Bayes algorithm

The sample space of test E is Ω. Suppose that "B1, B2,…, Bn" is a partition of Ω, A is an event of R, and P (A) > 0, P (Bi) > 0 (i = 1, 2, …, n), then the Bayesian formula is expressed as [6]:

$$P(Bi \mid A) \frac{P(Bi)P(A \mid Bi)}{\sum_{j=1}^{n} P(Bj)\, P(A \mid Bj)} \tag{6}$$

4 Experiment and Result Analysis

4.1 Experimental Environment

The programming language is python 3.6. The framework of deep learning is Tensorflow1.0. The internal storage is 8 GB. The operating system is windows 10.

4.2 Data Analysis

The wiki Chinese corpus used for text data contains 863000 Chinese texts. There are more than 80000 pieces of data belonging to spam SMS category, accounting for 10% of the total number of data. There are more than 720000 pieces of data belonging to

Table 3. The format of Chinese text data

Number	Category	SMS text content
1	0	"The secrecy of trade secret is one of the preconditions to maintain its commercial value and monopoly position."
2	1	Thank you for calling Hangzhou Xiaoshan Quanjin kettle Korean barbecue shop, located at XXX Jincheng Road. Korean barbecue, etc., affordable, welcome to patronize [Korean barbecue restaurant of Quanjin kettle]
3	0	Bring us a grand visual feast in Changzhou
4	0	There are unexplained urinary stones, etc.
5	0	Feel self weight loss, jump weight loss Aerobics

normal SMS category, accounting for 90% of the total number of data. The data format is shown in Table 3.

75% of the data is used as the training data set, and 25% of the data is used as the test data set after randomly scrambling the data. Then segmented training data to training various classifiers, and then the trained classifier is used to predict the test data. The prediction results are compared with the actual results, and the classification accuracy of the classifier is obtained.

The performance report is generated by B function in a, and the learning curve analysis is constructed when the model parameters are determined. The training speed is shown in Fig. 3.

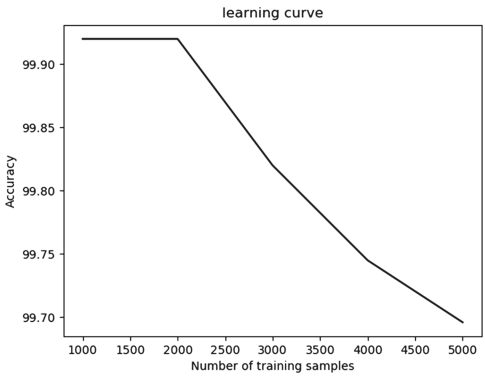

Fig. 3. Learning curve of naive Bayes classifier with prior Gaussian distribution

4.3 Interpretation of Result

The classification model trained by the training data set and the test data are used to predict the results. The performance reports generated are shown in Table 3 and Table 4. Because of the computational power 10074 data were selected for model training.

Table 4. Performance report of naive Bayes classifier 1

Category	Precision	Recall	f1_score	Support
Normal	93%	100%	96%	3925
Garbage	100%	29%	45%	412

Table 5. Performance report of naive Bayes classifier 2

Accuracy			93%	4337
Macro avg	97%	64%	70%	4337
Weighted avg	94%	93%	91%	4337

From the data in Table 4 and Table 5, it can be seen that 10074 data sets are used as raw data, and then the original data sets are divided into training data sets according to 75% of the comparison columns, and 25% of the comparison columns divide the original data sets into test data sets, the accuracy of naive Bayesian classifier reaches 93%. If the calculation force allows, the accuracy of naive Bayesian classifier can reach 93%, According to the actual situation, the training data set and the test data set data set data can be improved to some extent.

5 Conclusion

It is a social problem that spam SMS flooding has always plague people's life. In order to effectively identify spam messages, this paper proposes to use naive Bayes model to train samples from wiki Chinese corpus. To a certain extent, it solves the problems of sparse data, high dimension and difficult modeling of semantic relationship between words. However, when the number of features in the text is large or the correlation between the features is large, the classification effect has some limitations. At the same time, the prior probability needs to be known in the prediction, and the calculation of the prior probability depends on the assumptions of the model, which increases the difficulty of the prediction efficiency. Therefore, further research should be done to improve data classification and efficiency.

References

1. Bhat, S.Y., Abulaish, M.: Community-based features for identifying spammers in Online Social Networks. In: IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining (2013)
2. Baozhong, Z.: Application of various classification methods in garbage ports, pp. 5–45. Huazhong Normal University (2017)
3. Mountain, X.: Study on word vector classification, pp. 15-16. Jilin University (2019)

4. Ruiqi, C.: Research on the study of text acquisition analysis based on strengthening learning, pp. 11–12. Beijing University of Posts and Telecommunications, Beijing (2019)
5. Sone, S.F.: Spam identification application based on machine. Anhui Comput. Knowl. Technol. **16**(3), 202–204 (2020)
6. Huangyu, L.: Discussion on the application of Bayesian formula. Chengdu Sci. Technol. Econ. **28**(8), 165 (2020)