2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-1-6654-1364-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCWAMTIP53232.2021.9674063

SMS TEXT CLASSIFICATION MODEL BASED ON MACHINE LEARNING

XIAO FEI[1], LI JIANPING[1], GAO YUAN[1], ZHOU YUE[1]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
E-MAIL: 202021080416@std.uestc.edu.cn, jpl2222@uestc.edu.cn, 202021080403@std.uestc.edu.cn, 1026403954@qq.com

Abstract:

Text classification is an important problem in natural language processing. The main task is to divide the text into different categories according to the content of the text. This article preprocesses the text in the SMS data set used to a certain extent, using the Tf-Idf model. The frequency of the text unit is counted as the feature value of the corresponding vector of the text, so that the text is converted into a vector, and then these vectors are fitted and predicted by the support vector machine algorithm.

Keywords:

Machine Learning; Text Classification; Bag of Words

1. Introduction

Let computers understand human language is the main task of natural language processing, and the natural particularity of language makes this task difficult. Language is created by humans to express people's thoughts and feelings, whether in Chinese, English or other languages have their own grammars, but at the same time they have different expressions according to each language, each region and even each person's usage. These differences make the task of natural language processing particularly difficult. Fortunately, the enhancement of computer performance and the emergence of tools such as machine learning has allowed us to make great progress in this field.

In terms of text classification models, traditional word bag models and simple machine learning algorithms can be used, mainstream word embedding methods and deep learning algorithms can also be used, and the latest language models can also be used. Generally speaking, the more advanced algorithms can get better results, but usually they also need more time to compute. Therefore, each method has its application.

When the amount of data is sufficient and the text content is long and complex, deep learning and the latest language models can achieve better results. For shorter and simpler texts, traditional machine learning methods can achieve very good results. As a result, in comparison, the effect of using algorithms such as deep learning is far less than the price paid. This article will use traditional machine learning methods for the classification of harassment text messages. Since text messages are usually short and concise, traditional methods can achieve good results in a very short time. This article will use the SMS Spam Collection data set [1], combined with the improved bag-of-words model and support vector machine algorithm to classify the text in the data set.

2. Text preprocessing

Using traditional machine learning models in text classification, the preprocessing of text is a very critical step, and it can even be said to determine the quality of the model. The text content cannot be directly used for the training of the model. It is usually necessary to convert the text into a feature vector first. The most classic model is the Bag-of-Words model. The Bag-of-Words model firstly records all the words in the data set to form a vocabulary list, and then counts the frequency of words in each text in the vocabulary list, that is, each word in the vocabulary list is a feature, so each text can be represented by a vector, the dimension of the vector is the length of the vocabulary, this method is very simple, but also brings serious problems.

The first is the dimensionality of the vector. It can be imagined that when the data set is larger and the text content is richer, the vocabulary will become larger and larger, and the dimensionality of the vector representing the text will be larger, and each vector is very sparse, so it must get some pre-processing steps first. Secondly, the vector formed by counting the frequency of words will lose the spatial information of the text, thereby affecting the quality of the model. This problem can be alleviated by the N-grams method [2].

N-grams combine two adjacent words in a text into another unit. 2-grams is also called Bigrams, which uses two words as a unit to form a feature. Compared with a single word, it can provide a certain degree of text space information. Fig.1 show the difference between 1-grams and 2-grams on SMS spam Collection data set.

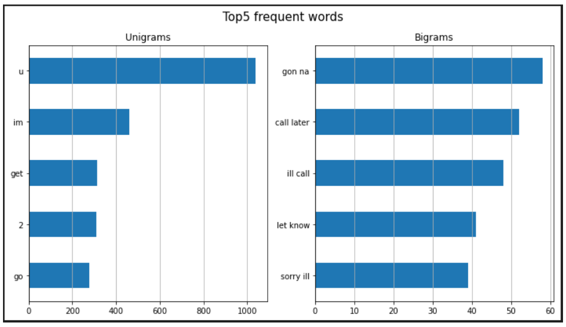

Fig.1 1-grams and 2-grams

In addition, the high frequency of words in the vocabulary does not necessarily mean that their contribution to the text is high. Some commonly used words usually appear very frequently, but the sentence information they provide is very scarce. This problem can be solved through the term frequency-inverse document frequency. Term frequency–inverse document frequency also called tf-idf is an advanced word bag model usage. The value of the feature vector of the ordinary word bag model is obtained by simple word count, while tf-idf comprehensively considers the frequency of the first word and the word The frequency of occurrence in all texts is obtained. We use the word frequency $tf_{i,j}$ to measure the importance of word i to the specific text j in which it is located, as the (1) shows.

$$tf_{i,j} = \frac{w_{i,j}}{\sum_k w_{k,j}} \tag{1}$$

Where the numerator $w_{i,j}$ represents the number of times the word i appears in the text j, and the denominator represents the total number of words in the text j.

It is one-sided to consider only the importance of a word to a specific text, and also consider the frequency of the word in all texts. If the word i appears in a large number in all texts, then the impact of the word on text classification is minimal. We use the reverse document frequency idf_i in (2) to measure the universality of word i for all texts.

$$idf_i = log \frac{|F|}{1+\left|\{j|w_i \in f_j\}\right|} \tag{2}$$

Where the numerator represents the total number of texts, and the numerator represents the number of texts containing the word w_i. Adding one can prevent the denominator from being zero.

The value of $tf_{i,j}$ represents the frequency of word i in text j, the value of idf_i restrains the power of word i, we use $tf - idf_{i,j}$ to represent the weight of the word w_i in the vector of text j by (3).

$$tf - idf_{i,j} = \; tf_{i,j} \; \times \; idf_i \tag{3}$$

2.1. Text segmentation and stop words

The text content is usually composed of a series of sentences, and a sentence is composed of several words in a certain order. To transform a text into a vector, the text needs to be segmented first, that is, each word of the text is segmented. In order to preserve the spatial information of the sentence, we will use the bi-grams method to combine single words in pairs to form a unit, and gather these units together to form a vocabulary.

The vocabulary at this time usually contains some words that have no effect on text classification such as the. Therefore, it is necessary to remove this stop word, and the vocabulary left after removal is used as the feature of the feature vector, so as to convert the text sentence 1 to sentence n is transformed into the vector 1 to vector n, and all the vectors are formed into a matrix, which is called the characteristic matrix.

2.2. Feature selection

The characteristic matrix obtained from the above is a sparse matrix. The direct use of the matrix is not good and adds a lot of time overhead. Therefore, we need to reduce the dimensionality of it, and use a binary bit to represent the category of each text. For the feature vector Perform a chi-square test with the corresponding category, and only retain the features that meet the specific p-value, so that the dimension of the feature matrix is greatly reduced.

The feature matrix at this time retains most of the statistically significant information, and has a smaller dimension compared to the original information, which can be used for the training of the classification model.

3. Text classification model

3.1. Split Dataset

The SMS Spam Collection data set used in this article contains 5572 short messages and has been marked as harassing short messages. Ham means normal short messages and spam means harassing short messages. The comparison of the data volume of each type of short message is shown in Fig.2.

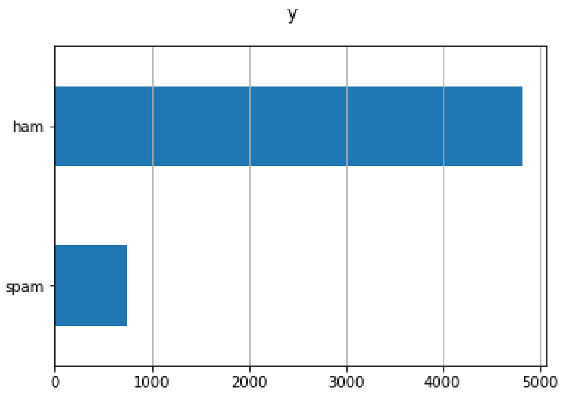

Fig.2 Dataset Structure

We divide the entire data set into a training set and a test set at a ratio of 7:3. We will train the model on the training set, test our model on the test set and give the model's effect evaluation.

3.1. Support Vector Machines

Support vector machine [3-4] is the most popular tool in traditional machine learning. Before deep learning became popular, it was the mainstream method to solve various problems. Although its model effect is no longer comparable to other advanced learning methods, in some aspects, it is still Has an irreplaceable role. In this article, we will use the support vector machine algorithm to train the preprocessed feature matrix, and then make predictions on the test data.

The support vector machine finds a hyperplane to segment data of different categories as much as possible, (4) shows how to decide vector's class.

$$class = sign(\theta^T \cdot w_i + b) \tag{4}$$

Where θ and b are the parameters obtained by optimizing the following objective function

$$\min_{\theta} \left\{ \Sigma_i^m \left[y^i \cdot cost_1\left(\theta^T \cdot x_i\right) + \ (1 - \ y^i) \cdot cost_0\left(\theta^T \cdot x_i\right) \right] + \ \frac{\lambda}{2} \|w\|^2 \right\} \tag{5}$$

Where $cost_1$ and $cost_2$ represent different types of cost functions

$$cost_1 = -\log \frac{1}{1+ \ e^{-\theta^T \cdot x}} \tag{6}$$

$$cost_0 = \ -\log\left(1 - \ \frac{1}{1+ \ e^{-\theta^T \cdot x}}\right) \tag{7}$$

The final parameters can be optimized by methods such as gradient descent, and then the trained model can be used to predict the results.

4. Experimental results

There are many indicators for evaluating a model. We selected the most common accuracy, recall, and F-1 values. The specific values are shown in the Table 1.

Table 1 Confusion Matrix

	precision	recall	f1-score
Ham	0.99	0.98	0.99
Spam	0.90	0.92	0.91

It can be seen that our model has achieved good results on this data set. Only from the data point of view, it is more abstract and can be more intuitively understood through some visualization tools. Confusion matrix is a common visualization tool. The confusion matrix of the model is shown in the Fig.3.

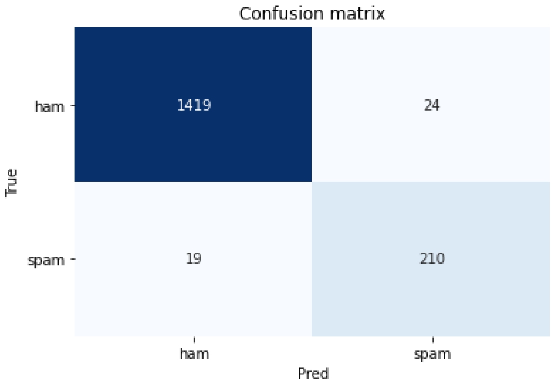

Fig.3 Confusion Matrix

The other two commonly used visualization tools are ROC curve and PR curve. Fig.4 has shown the ROC curve and PR curve of our model.

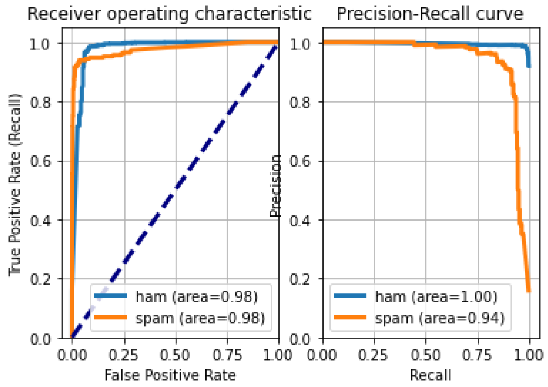

Fig.4 ROC curve and PR rate curve

Through these visualization tools, we can feel the effect of the model more intuitively.

5. Conclusions

Text classification is the most common and very important type of task in natural language processing. The algorithms and models that can be used for text classification are also diverse. For data sets such as harassment text messages used in this article, traditional machine learning algorithms can be used. Achieve good results, but also does not need to spend a high computational cost, therefore, in the case of complex text content, traditional machine learning methods are also useful.

Acknowledgements

This work was supported by the National Natural Science Foundation of China, the National High Technology Research and Development Program of China, the project of Science and Technology Department of Sichuan Province(Grant No. 2021YFG0322), the project of Science and Technology Department of Chongqing Municipality, the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202114401)，Chongqing Qinchengxing Technology Co., Ltd., Chengdu Haitian Digital Technology Co., Ltd., Chengdu Chengdian Network Technology Co., Ltd., Chengdu Civil-military Integration Project Management Co., Ltd., and Sichuan Yin Ten Gu Technology Co., Ltd.

References

[1] Almeida, T.A., Gómez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

[2] Zečević, A. N-gram based text classification according to authorship. In Proceedings of the Second Student Research Workshop associated with RANLP 2011, pp. 145-149, September 2011.

[3] Colas, F., & Brazdil, P. Comparison of SVM and some older classification algorithms in text classification tasks. In IFIP International Conference on Artificial Intelligence in Theory and Practice, pp.169-178. Springer, Boston, MA, August 2006.

[4] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974, 2005.