An Exploration on Text Classification with Classical Machine Learning Algorithm

Yuhan Zheng
Southwest Petroleum University, Chengdu, Sichuan, China
rockzhengyuhan@163.com

***Abstract*. Text classification is an essential part of the NLP, which aims to predict the categories for given texts in a particular classification system. There are many ways of feature selection and classification models. However, most researchers would like to use the encapsulated methods of third-party libraries to achieve their goals. Therefore, in this paper, we propose to implement code to achieve functions, instead of using third-party libraries. We evaluate our code in different classification models, and the result of our experiment shows that our code is feasible.**

Keywords: text classification, machine learning, natural language processing, text classification

I. Introduction

Along with the fast development of the Internet, the amount of information on the network is growing exponentially. There are various carriers of information, such as pictures, video, audio, and texts. To process such a large number of information, a verity of academic fields have been derived. NLP (Natural Language Processing) is an important area in computer science and artificial intelligence [1]. Besides, text classification has been one of the important parts of NLP.

Text classifications have made great performances in the NLP area. In the past few years, there were a lot of previous works about text classifications [2], [3]. Feature selection is one of the important parts of the text classifications., Rajni Jindal, Ruchika Malhotra and Abha Jain summarized them in their research, and they also compared the different techniques in data mining and document representation [4]. Moreover, M.Thangaraj and M.Sivakami introduced some standard text classification techniques and compared their performance in their study [5].

Recent years, more and more people have realized the importance of text classification. For example, text classification can help us work more efficiently and save a lot of times such as library classification, news classification, and so on. Therefore, there are much more applications of text classification in specific areas. Yang and her team designed a model of text classification to solve the problem in courts' documents, which are lack of automatic management in the process of informatization [6]. Jin and Song designed a model of emotion text classification for E-commerce comments, and it compared three classification models and SVM [7]. In other fields, such as language learning, Tulnisa Guri semiti designed a model of text classification for learning Uygur[8]. Some researchers devote themselves to improves and tests the models [9]. Alternatively, researchers put their efforts on classification techniques and analyze its advantages and disadvantages [10], [11].

Thanks to these researches on text classification, there are much more applications of text classification in our life. Such as news classification [12], space knowledge management [13], emotional analysis [14], and commodity evaluation [15]. However, in most of these studies, researchers would like to use the encapsulated methods of third-party libraries to achieve their goals, such as building the Bag-of words, one-hot encoding and so on.

The significant contributions of this paper are summarized as follows:

What related works the researchers did in text classification, and the work we did in this experiment. Besides, the introduction of the models we used in this experiment.

This article summarized below. In Section 2, we present some previous work on the text classification task. We introduce the classification models which we use in our experiment in Section 3. Section 4 details that we compare the performance of our models. Section 5 is what we found in our experiments. The last Section is our conclusion and things we will do in the future.

II. Related Work

Our research in text classification includes:

1.Split the words and remove the stop words.

2.Feature selection, show different characteristic forms of words such as One-hot, Bag-of-word, TF-IDF.

3.Evaluate the performance of different methods on our dataset.

In our experiment, we did not use third-party encapsulated libraries or methods to make functions. However, we implemented the vectorization module, which includes one-hot encoding, bag-of-words and, TF-IDF, and so on. Liao and Yan "use the methods in Sklearn library to calculate TF-IDF [16]." Rajini Jindal and her team "compare and analyze the different feature selection methods(such as TF, DF, ARM, CMFS)in their study[4]."In the study of M.Thangaraj and M.Sivakami "Introduce the different models of text classifications, and compare them [5].".We read many types of research to learn about text classification, such as Huang's

978-1-7281-5094-9/19/$31.00 ©2019 IEEE
DOI 10.1109/MLBDBI48998.2019.00023

"research in Chinese text classification [17]."

One-hot encoding is not only useful in text classification. In other studies, Liang "use one-hot encoding in the neural network [18]." The normalization is also necessary for text classifications; standardization aims to reduce errors in the data. We learn a lot about normalization in Zhang and Chen 's study," normalization will solve the problem, which is lack of parameter estimation in Naïve Bayes [19]." We also learn a lot about the algorithm in text classification in Yao's study [20]. According to the study of the above experiments, we found that most researchers would choose TF-IDF as a way to select the feature. In the selection of text classification models, Naïve Bayes Classifier and Logistic regression are also popular. The mechanism of these methods is simple, and effect to text classification is useful.

The K-Nearest Neighbor classifier (KNN), which decide a document i whether belongs to specific class according to nearest samples, is used for text classification [21]. K-NN sorts the unlabeled sample with the same class for the largest proportion of neighbors which is assigned to class j. Decision trees [22], which are widely used in inductive learning and classification task, has capability to carry out suitable document classification. classification. One of the most famous decision tree algorithms is ID3, C4.5 and C5, which is a top-down method to build a classifier. Then A Support Vector Machine (SVM) [23][24] is a classification algorithm that has been successfully used for text classification and other classification tasks, while Logistic Regression (LR) were fit to experimental data for binary text classification [25].

III. Model

In this section, we will introduce the models we used in this experiment.

A. TF

Term Frequency evaluates how frequently a term occurs in a document. The formula as follow:

$$\mathrm{TF} = \text{Number of times term t appears in a document} \tag{1}$$

Since every document is different in length, and no words are unique in a document. Thus, the term frequency is often divided by the document length as a way of normalization:

$$\mathrm{TF} = \frac{(\text{Number of times term t appears in a document})}{(\text{Total number of terms in the document})} \tag{2}$$

B. IDF

Inverse Document Frequency evaluates how important a term is. It means the number of documents which include the term. More documents include it, more ordinary the word is.

$$\mathrm{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term in it}+1}\right) \tag{3}$$

If a word is common, the denominator will be larger, and the IDF will be smaller. The reason for adding 1 in the denominator is to avoid 0 (all documents do not contain the word).

C. TF-IDF

Measure the value of the attribute for each word in documents.

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} \tag{4}$$

It is directly proportional to the number of occurrences of a word in a document and inversely proportional to the number of events of the word in the whole corpus.

D. Naïve Bayes Model

The core algorithm of Naïve Bayes is:

$$\mathrm{P}(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{5}$$

$P(B|A)$ means the probability of event B occurring on the premise that the event A has occurred (conditional probability). The formula as follows:

$$\mathrm{P}(B|A) = \frac{P(BA)}{P(A)} \tag{6}$$

$$\mathrm{P}(A|B) = \frac{P(AB)}{P(B)} \tag{7}$$

After a simple transformation:

$$\mathrm{P}(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{8}$$

The whole Naive Bayesian classification is divided into three stages:

Preparatory stage; The main work is determining the feature attributes according to the specific situation, and share each feature attribute appropriately, and then manually classify a part to be organized to form a training sample set.

Classifier training stage; In this stage, the task is to generate a classifier. The main work in this stage is to calculate the occurrence frequency of each category in the training sample and the conditional probability estimation of each group divided by each characteristic attribute and records the result. The input is the feature attribute and training sample and the output is classifier.

Application stage; The task of this stage is to classify the classification items by using a classifier, whose input is the classifier and the things to be organized, and the output is the mapping relationship between the items to be classified and the categories.

E. Logistic Regression Model:

The definition of logistic regression is: If X is a continuous random variable, and X obeys the logistic distribution, it means that X has the following distribution functions and density functions:

$$F(X) = P(X \le x) = \frac{1}{1+e^{-\frac{(x-\mu)}{\gamma}}} \tag{9}$$

$$\mathrm{f}(x) = F'_{(x)} = \frac{e^{-(x-\mu)/\gamma}}{\gamma\left(1+e^{-(x-\mu)/\gamma}\right)^2} \tag{10}$$

μ is the position parameter, $\gamma > 0$ is the shape parameter.

Figure 1: Density function and Distribution function

As Figure 1, we can see the Density function and Distribution function of Logistic Regression model. The Logistic regression model is a classification model, expressed by a conditional probability distribution, in the form of the parameterized logistic distribution. The random variable is the real number and is 1 or 0. The formula as follow:

$$P(Y=1|x)=\frac{exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \quad (11)$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)} \quad (12)$$

Here, $x \in R^n$ is input, $Y \in \{0,1\}$ is output, $w \in R^n$ and $b \in R$ are trainable parameters, w is weight vector, b is bias, $w \cdot x$ means the inner product of w and x.

From another view, we consider the liner function, which classifies the input, and its range is the real number. The linear function can be transformed into probability by logistic regression model definition. The formula as follow:

$$P(Y=1|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)} \quad (13)$$

At this time, the closer the value of the linear function is to positive infinity, the closer the probability value is to 1; the closer the amount of linear function is too negative infinity, the closer the probability value is to 0.

F. SVM:

Support Vector Machine (SVM) is a binary classification model. The basic model is: the linear classifier with the most significant spacing in feature space is defined, which makes it different from perceptron. The kernel technique, which makes it essentially a non-linear classifier, is the key technology to SVM. The learning algorithm of support vector machine is the optimal algorithm for solving convex quadratic programming. Given a linear separable training data set, the separable hyperplanes obtained by maximizing or equivalently solving the corresponding convex quadratic programming problems are as follow:

$$w^* \cdot x + b^* = 0 \quad (14)$$

Classification decision function:

$$f(x) = \text{sign}(w^* + b^*) \quad (15)$$

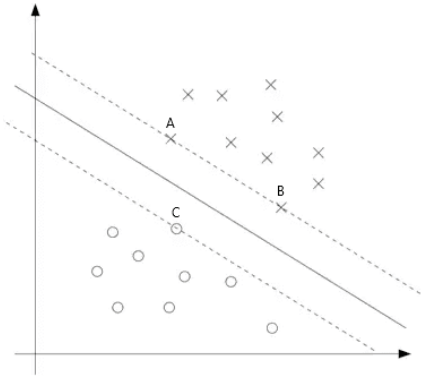


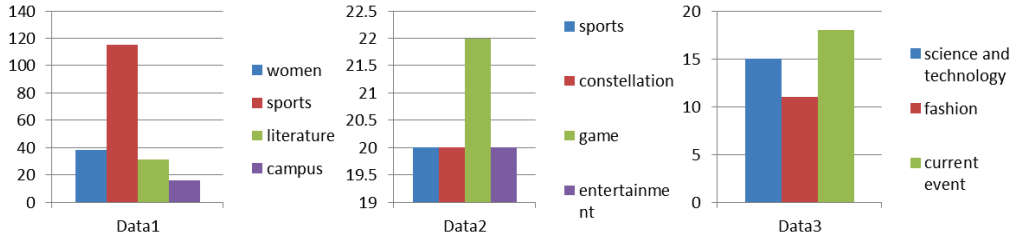
Figure 2:The geometric representation of a hyperplane

As shown in figure 2, " ○ "are negative examples, " × " are positive examples. The aim of SVM is finding a hyperplane which could solve the problem of binary classification. We first find the nearest point of each classification sample to the hyperplane, so that the distance between the point and the hyperplane can be maximized. The marks on the dotted line are the nearest. These points (A, B, C) can be used to determine a hyperplane. And it is also a vector expressed in geometric space, so these vectors which can be used to determine the hyperplane are called support vectors.

IV. Experiment

A. Datasets

We used three datasets for testing, and the first test set (Data1) has four categories, it includes women (38), sports (115), literature (31) and campus(16).The second test set (Data2) also has four categories, sports (20), constellation (20), game (22), entertainment (20). The third test set (Data3) has three categories, includes Science and Technology (15), fashion (11), current event (18). We collect the samples from news websites. Most of them are Chinese, but some of them are mixed with special symbols and URLs. The Figure 3 shows the data in each dataset:



(a) Data in dataset1 (b) Data in dataset2 (c) Data in dataset3

Figure 3:Data in each dataset

B. Evaluation Metrics

Before explaining the evaluation index of text classification, we first draw up four categories:

－TP- predict positive class as positive class (correct prediction)

－FN- predict positive class as negative class (wrong prediction)

－FP- predict negative class as positive class (wrong prediction)

－TN- predict negative class as negative class (correct prediction)

- Accuracy

$$Accuract = \frac{Correctly\ predicted\ sample}{Total\ sample} = \frac{TP+TN}{TP+FN+FP+TN} \quad (16)$$

- Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

- Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad (18)$$

- F1 value

$$F1\ value = \frac{\text{Accuracy} \times Recall \times 2}{(\text{Accuracy} + Recall)} \quad (19)$$

C. Experiment Settings

In our experiment, the models of text classification we use are Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Logistic Regression CV (LRCV). We compare the results in different n-gram setting (1, 1) and (1, 2).

Table 1:n-gram is (1,1)

Model	Data1			Data2			Data3		
	Precision	Recall	F1value	Precision	Recall	F1value	Precision	Recall	F1value
NB	0.89	**0.87**	**0.85**	0.72	0.67	0.67	0.63	0.54	0.51
SVM	0.65	0.30	0.28	0.27	0.29	0.18	0.12	0.33	0.18
LR	**0.94**	0.81	**0.85**	0.77	0.74	0.74	**0.72**	0.42	0.35
LRCV	0.85	0.83	0.83	**0.84**	**0.77**	**0.78**	0.64	**0.63**	**0.63**

Table 2:n-gram is (1.2)

Model	Data1			Data2			Data3		
	Precision	Recall	F1value	Precision	Recall	F1value	Precision	Recall	F1value
NB	0.88	**0.87**	**0.86**	0.76	0.71	0.71	0.67	**0.57**	**0.56**
SVM	0.40	0.28	0.24	0.27	0.29	0.18	0.12	0.33	0.18
LR	**0.94**	0.74	0.79	0.76	0.71	0.71	**0.70**	0.40	0.32
LRCV	0.83	0.85	0.83	**0.81**	**0.76**	**0.77**	0.57	0.56	0.55

D. Result and Discussion

By the Table 1, when n-gram is (1, 1), we could find that LRCV has a better performance. In the classification of Data 2, it has the highest Precision, Recall, and F1 value, about 0.84, 0.77 and 0.78. In the classification of Data 3, it also has the highest Recall, and F1 value, both of them are 0.63. The second is the LR. In the classification of Test, it has the highest Precision and F1 value, 0.94 and 0.85. SVM is a widespread text classification model, but it does not have excellent results in this classification.

In Table 2, when we change n-gram to (1, 2), it can be seen that there is a significant decrease in overall results. There have also been some changes in the performance of the classification model.

NB (Naïve Bayes) becomes the model which has the best performance. In the classification of Data 1, it has the highest Recall and F1 value (0.87 and 0.86). In the classification of Data 3, it also has the highest Recall and F1 value, which are 0.57 and 0.56. The second is LRCV. In the classification of Data 2, the three items are most top, about 0.81, 0.67, and 0.77 respectively. We can see that the LRCV model has a better performance in text classification of Data 2. It means that when the dataset is small and average, we can choose LRCV as our model of text classification.

V. conclusion and future work

In conclusion, the selection of n-gram can affect the result of text classification to some extent. We should choose the right model to do the classification when the texts are different.

Although we did not implement some methods by using third-party libraries, we did it by implementing our code. Moreover, the results show that the code we implement can deal with some small text classification. However, there are still many problems in text classification at present. The machines still need to be trained al lot by a large number of training sets, which will take much time. Also, the selection of a classification model is also different in different cases, and there is no classification model which is suitable for each case. In future work, we will give more consideration to the selection of classification models and the training of machines. The machines still need to be trained al lot by a large number of training sets.

Acknowledgements

First of all, I would like to thanks my mentor, Ph.D. Pan Disheng. Without his patient and careful guidance, this paper cannot be finished. He gives me much help no matter in this paper or in my study. I also want to thanks my parents for their supporting. Last but not least, I have to thank all the friends for their encouragements.

Reference

[1] Zhao J S, Song M X, Gao X. Review of the development and application of NLP[J]. Information technology and informatization,2019(07):142-145

[2] Wang Y. Text classification and its application based on Natural Language Processing and Machine Learning[D]. Graduate school of Chinese Academy of Sciences (Chengdu institute of computer application),2006

[3] Michael D. Lee,Elissa Y. Corlett. Sequential sampling models of human text classification[J]. Cognitive Science,2003,27(2).

[4] Rajin Jindal, Ruchika Malhotra, Abha Jain. Techniques for text classification: Literature review and current trends.

[5] M.Thangaraj, M.Sivakami.Text classification techniques: a literature review

[6] Yang F, Chen J X, Zheng Y Q, Huang Y J, Li C. Classification of court information texts based on deep learning [J]. Journal of Hubei university of technology,2019,34(04):63-67.

[7] Jin L, Song W G. Text emotion classification based on e-commerce comments[J]. Computer knowledge and technology,2019,15(11):290-292.

[8] Tulnisa guri semiti. Classification research and systematic realization of uygur texts based on n-gram [D]. Xinjiang university,2014.

[9] Chen J N, Dai Z B, Duan J T,Heinrich Matzinger,lonel Popescu. Improved Naive Bayes with optimal correlation factor for text classification. School of Mathematics, Georgia Institute of Technology. SN Applied Sciences, 2019, Vol.1 (9), pp.1-10

[10] Li W Q, Wei L, Jia C S. Text classification based on logistic model [J].China high-tech zone,2018(03):31-32.

[11] Shi C Y, Xu C J, Yang X J. Review of TFIDF algorithm [J].Computer application,2009,29(S1):167-170+180.

[12] Yan S. Research on text classification oriented to news [J]. Computer knowledge and technology,2019,15(16):283-284.

[13] Guo S. Research on information collection and classification application of knowledge management in the field of aerospace [D].National space science center,Chinese Academy of Sciences,2016.

[14] Luo Y P. An analysis and research on the emotional tendency of Chinese comment text oriented to online public opinion [D]. Northeast university of finance and economics,2010.

[15] Zhong J, Yang S Y, Sun Q G. Emotional analysis of commodity evaluation based on text classification [J]. Computer application,2014,34(08):2317-2321.

[16] Liao Y X, Yan S R. Implementation of Chinese text classification based on Python [J]. Fujian computer,2016,32(12):6+14.

[17] Huang X.Research and implementation of Chinese text classification based on machine learning [D]. Heilongjiang university,2016.

[18] Liang J, Chen J H, Zhang X Q, Zhou Y, Lin J Q. Anomaly detection based on on-hot and convolutional neural network [J].Journal of Tsinghua university (natural science edition),2019,59(07):523-529.

[19] Zhang J, Chen H X.Text classification method based on normalized word frequency bayesian model [J]. Computer engineering and design,2016,37(03):799-802.

[20] Yao F. Research on Chinese text classification based on python [D]. Huazhong university of science and technology,2016.

[21] Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys. Vol 34, pp. 1 –47.

[22] Lewis, D. D., and Ringuette, M. 1998. A comparison of two learning algorithms for text classification. In the Proceedings of Third annual symposium on Document Analysis and Information Retrieval, pp. 81–93.

[23] Hotho, A., Nürnberger, A., and Paaß, G. 2005. A Brief Survey of Text Mining. Journal for Computational Linguistics and Language Technology. Vol. 20, pp. 19 – 62.

[24] Rigutini, L. 2004. Automatic Text Processing: Machine Learning Techniques. Ph.D. Thesis, University of Siena.

[25] Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, 9(4), 705-724.