A Study of Intelligent Reading Model

Yu Jun
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, ShaanXi, China
e-mail: yujun@xatu.edu.cn

Hu Zhiyi
Engineering Design Institute
Army Academy of PLA
Beijing, 100000, China
e-mail: huzhiyi016v7@163.com

Kang Qinyu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, ShaanXi, China
e-mail: 534739457@qq.com

Li Zhonghua
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, ShaanXi, China
e-mail: 761173763@qq.com

***Abstract*—In order to solve the problem of how to find out the required information quickly from a large number of reading texts, this paper constructs an intelligent reading model. The model adopts the principle of "the minority obeys the majority". The results of the classifier trained by the three algorithms, those are decision tree, Bagging and Gauss Bayes algorithm, are filtered to build an intelligent reading model. Based on the experimental results, the objective evaluation results of the new combinatorial algorithm are attained.**

Keywords-Natural language processing(NLP); Decision Tree; Bagging; Gaussian Bayesian Algorithm

I. INTRODUCTION

In recent years, with the rapid developing of the Internet and other emerging media, human beings have entered the era of information explosion. At the same time, more and more people hope that computers can understand human language so as to help human beings to perform various daily tasks better. Natural Language Processing(NLP)[1], as a typical example of artificial intelligence application in the practical field, is a necessary means for modern people to mine a large amount of data and information. Its main goal is to let computers learn to understand and use human natural language. Therefore, Natural Language Processing (NLP) has become a research hot spot in recent years.

At present, as one of the representative products of Natural Language Processing(NLP), "smart interactive technology [2]" has gradually penetrated into many products. However, many smart products can only recognize some specific commands. For example, when the input is "Open QQ (QQ is the abbreviation of Tencent QQ, which is an Internet-based instant messaging software developed by Tencent Company)", it can start QQ. But the input is "Look at QQ" and nothing happens. In addition, people have to read a lot of texts in daily life, such as novels, tutorials, etc. Sometimes you can solve the problem by just looking for a small part of the text without having to read through the whole article. For example, we can solve our legal doubts by looking for certain passages in the legal literature and be unnecessary to read the entire legal literature. Based on this, in order to make our reading more "intelligent", we need to establish an intelligent reading model that can use natural language to communicate with machines and let machines serve us in order to minimize the learning burden.

This paper builds an intelligent reading model. Since English is based on words, which are separated by spaces. However, Chinese is in the form of word, each of words in one Chinese sentence have to be connected for describing a complete meaning. For example, the English sentence "I am a student", Chinese means "I am a student". In English, the computer can easily know that "student" is a word by Spaces, However, in Chinese, "student" is made up of two words, which can only be combined to mean a word. Therefore, Chinese word segmentation is to divide the sequence of Chinese characters into meaningful words. Due to certain uncertainty in Chinese word segmentation, it is necessary to adopt many different technologies such as Jibe word segmentation [3] and TF-IDF weight algorithm [4]. In view of the simpleness of the previous model, a new combination model, namely our intelligent reading model, is constructed by combining multiple algorithms as well as adopting the principle of "the minority obeys the majority". Among them, the principle of "the minority obeys the majority" means that if a data is trained by three classifiers, the output result of classification is "0 / 0 / 1", Because there is a large number of 0 in results, so the data of the final result is 0. Finally, the validity of the model is verified by experiment and calculation.

II. THE OVERALL PROCESS OF BUILDING AN INTELLIGENT READING MODEL

The establishment of intelligent reading model includes five parts, these are data acquisition, data processing, feature extraction, training classifier and building model. The overall process is shown in figure 1. The details are as follows.

DOI: 10.21307/ijanmc-2019-013

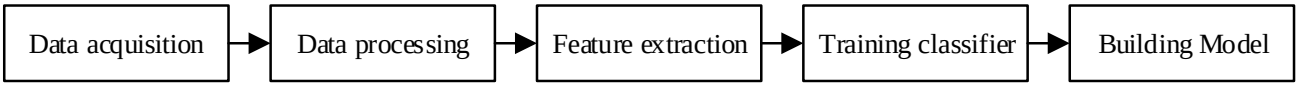


Figure 1. Overall framework

1) Data acquisition：Because there are all kinds of text on the network, and the quantity is huge, so the Python network crawler [5] is used to obtain the data. that is stored in a TXT file.

2) Data processing：Due to the inaccuracy of the word segmentation, it is necessary to use the algorithm based on information entropy to find new words, as well as put the new words into the custom dictionary, followed that process the text by Chinese word segmentation, stop word filtering and so on.

3) Feature extraction：For the processed data in step (2), the TF-IDF algorithm is used to extract the feature value and generate the word-text matrix.

4) Training classifier：Decision tree [6], Bagging [7] and Gaussian Bayesian algorithms [8] are used to train the word-text matrix generated by step (3).so that three classifiers could be obtained.

5) Building Model. For the three classifiers obtained in step (4), the principle of "the minority obeys the majority" is adopted to establish the intelligent reading model.

III. DESCRIPTION OF THE PROCESS OF MODELING

This paper constructs an intelligent reading model. Firstly, the Python network crawler is used for data acquisition. followed that Jieba word segmentation technology and TF-IDF weight algorithm were adopted to preprocess the sample data. Finally, Extracting the feature value, training classifier, establishing model and carrying out other operations. The detailed operation is described as follows.

A. Data acquisition

There are a variety of texts on the web. Due to the large number of data, Python web crawlers are usually used to obtain data. But some websites have anti-crawler mechanisms. Therefore, while designing a web crawler, a simulation operation of browser accessing is necessary. Through analyzing the web page source codes, regular expressions are used to obtain the required data. The library files, such as Beautifulsoup, Requests and Re can be used to crawl data. The content of the crawl is the problems and all their corresponding answers. Finally, the acquired data is stored in a TXT format file and is added a fixed tag name, so as to be convenient for later data processing.

B. Data processing

By analyzing the acquired data, a lot of noise in the text information can be found. For example, word segmentation is inaccurate, as well as there are a large number of stop words. If these noises are brought into the operation of word frequency statistics, it will not only reduce the processing speed, but also greatly affect the experimental results. Therefore, the first important thing is to preprocess the data. Data preprocessing is divided into three steps, which are generating and loading of the custom dictionaries, Chinese word segmentation and stop-word filtering. It is as shown in figure 2.

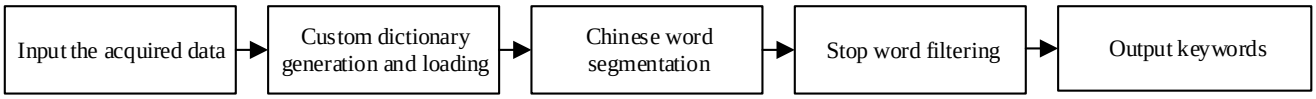


Figure 2. Flow of data preprocessing

1) Generation and loading of Custom Dictionary

Due to the number of words included in the Jieba dictionary is limited, which leads to the inaccuracy of text segmentation. For instance, it is inaccurate segmentation of people's names and place names. So that it is necessary to require a custom dictionary to improve the accuracy of word segmentation. The information entropy algorithm is used to find new words and generate custom dictionaries. After that, the generated custom dictionaries are loaded into the codes to improve the precision of word segmentation.

2) Chinese word segmentation.

After the above steps are completed, the word segmentation is beginning. This paper adopts a Chinese word segmentation module developed by Python-jieba word segmentation, which divides all data sets in Chinese. It combines rule-based and statistics-based methods [9].

The rule-based method means that, the word segmentation based on an existing dictionary which adopts manual rules such as forward maximum matching, backward maximum matching and bidirectional maximum matching. For example, for the sentence "Shanghai tap water comes from sea", forward maximum matching is used. It scans from forward to back, as well as makes the separated words exist in the dictionary and lets the words as long as possible. At last the sentence of "Shanghai / tap water / from / sea" can be obtained. This kind of method is simple and easy to implement, and the required data amount is not high.

The statistics-based method is to summarize the probability distribution of words and the common collocation between words from a large number of manually labeled corpus, and supervised learning is used to train the word segmentation model. For the sentence "Shanghai tap water comes from the sea", the most basic participle method

based on statistics is to try all possible participle schemes, Because of any two words, or need to cut, or without segmentation. For all possible word segmentation schemes, the probability of each scheme is counted according to the corpus, and then the one with the greatest probability is retained. Obviously, "Shanghai / tap water / come from / sea" is more likely than "Shanghai tap / water / come from / sea ", Because "Shanghai" and "tap water" appear more frequently in tagged corpus than "Shanghai tap" and "water". The result of partial participle is shown in figure 3.

100001	question: lane,occupy,Emergency,highway,How many	10000101	content: lane,occupy,Emergency,How many scores,Driving
100001	question: lane,occupy,Emergency,highway,How many	10000102	content: lane,driver,Emergency,confusion,highway
100001	question: lane,occupy,Emergency,highway,How many	10000103	content: lane,occupy,Emergency,Deduction,Illegal
100001	question: lane,occupy,Emergency,highway,How many	10000104	content: lane,parking,Emergency,The new traffic rules,occur
100001	question: lane,occupy,Emergency,highway,How many	10000105	content: The road traffic,Regulations,legislation,¥200,¥20
100001	question: lane,occupy,Emergency,highway,How many	10000106	content: lane,Emergency,2017,occupy,highway
100001	question: lane,occupy,Emergency,highway,How many	10000107	content: Emergency situations,lane,parking,Emergency,The new traffic rules
100001	question: lane,occupy,Emergency,highway,How many	10000108	content: lane,Emergency,occupy,2017,punishment
100002	question: The king of kung fu,Bride with white hair,play	10000201	content: The king of kung fu,Bride with white hair,Jet li,Jackie Chan
100002	question: The king of kung fu,Bride with white hair,play	10000202	content: Bride with white hair,kung fu,Jet li,Mars,The film

Figure 3. Participle screenshot

In figure 3, it can be found that there are a large number of meaningless modal particle in the results of word segmentation, which will greatly influence the final experimental results. Therefore, it is necessary to carry out the filtering of stop words.

3) Filter stop words

Stop Words [10] refers to words that appear frequently in the text without practical significance. Such as modal particle, adverbs, prepositions, conjunctions, etc. In order to save storage space and improve search efficiency, these meaningless stop words must be filtered out before processing text. To find out the stop word accurately, the following indicators can be used to measure the effectiveness of words.

a) Term Frequency. TF is a simple evaluation function whose value is the number of words occurring in the training set. The theoretical assumption of the TF evaluation function is that, when one word appears frequently in the text, it is generally regard as a noise word.

b) Document frequency. Similar to Term Frequency (TF), the theoretical assumption is that when one word appears frequently in the text, the word is generally regard as a noise word. The experimental result is shown as Table 1.

TABLE I. STOP WORD LIST(PARTIAL)

category	Stop Words
preposition	On, In, At, under, Beside, Behind, To, Over, with
pronoun	Everyone, everything, everywhere
...	...
adverbs	So, still, therefore, moreover, however

As shown in table 1, using filtered stop words to generate stop-word list, as well as to load the list into codes. Followed that the result of word is matched with the words in the stop-word list. If the matching is successful, the word from the result of the segmentation will be deleted.

C. *Feature extraction*

After the preprocessing of the above steps, although the stop word is removed, the sentence still contains a large number of words, which brings difficulties to the text vectorization process. Therefore, the main purpose of feature extraction is to minimize the number of words for being processed without changing the core content of the original text, so as to reducing the dimension of vector space, simplifying calculation and improving the speed and efficiency of text processing. Commonly used methods include term frequency-inverse document frequency(TF-IDF), information gain [11], X2 statistics, etc. Hereby TF-IDF algorithm is used to transform keyword information into weight vector in here. The steps are described as follows.

1) Calculating the word frequency, which is TF weight.

$$TF = \frac{\text{The number of times a word appears in the text}}{\text{The total number of words in the text}} \quad (1)$$

2) Calculating the inverse document frequency. that is IDF weight.

Firstly, a corpus is required to build up for simulating the language environment. The larger the IDF, the more concentrated this feature is in the text, it means that the more able the words are to distinguish the content of the text.

$$IDF = \log\left(\frac{\text{Total number of texts in a corpus}}{\text{Number of text containing the word} + 1}\right) \quad (2)$$

3) Calculating the Term Frequency Inverse Document Frequency (TF - IDF) values.

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

The larger the TF-IDF value, the more important the word is. Calculating and sorting the TF-IDF value of each word in the text. The first six keywords of each question and corresponding answer in the text are found in turn, and the corresponding weight of the six keywords is returned. If there are less than 6 keywords, the residual weight is set to 0.

By using the TF-IDF algorithm, the text information is vectorized and the lexical text matrix is obtained. The details are described as follows.

$$\begin{array}{c} \quad d_1 \quad d_2 \quad K \quad d_n \\ \begin{array}{c} t_1 \\ t_2 \\ M \\ t_m \end{array} \begin{bmatrix} w_{11} & w_{12} & K & w_{1n} \\ w_{21} & w_{22} & K & w_{2n} \\ K & K & O & K \\ w_{m1} & w_{m2} & K & w_{mn} \end{bmatrix} \end{array}$$

Hereby, t_i (i=1,2,3..., n) is the feature item in document D. as well as w_{ij} (i, j=1,2,3..., n) is the weight of the feature item. The calculation formula is described as follows.

$$W_{ij} = TF\text{-}IDF = TF \times IDF \tag{4}$$

D. Training classifier

In this paper, the lexical - text matrix is trained with three algorithms of decision tree, Bagging and Gaussian Bayes. The specific steps of each algorithm are described below.

1) decision tree

Decision tree algorithm mainly includes feature selection and decision tree generation. The feature selection is based on the relationship between the information gain and the data set. According to the characteristics of the selected data set, the decision tree is generated recursively using ID3 algorithm [12]. The specific steps are described as follows.

a) Calculating information entropy.

In order to select the feature of good classification ability for training data, the information gain is introduced. And then, the calculation formula of information entropy is described as follows. Assume D is the training element group in the training set, its entropy can be expressed as follows.

$$info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{5}$$

$$p_i = \frac{\text{Number of elements in this category}}{\text{Total number of training tuples}} \tag{6}$$

Hereby, m represents the total number of categories, and p_i represents the occurring probability of Category i which appears in the entire training tuple.

Entropy is a measure of the uncertainty of random variables. The actual meaning is the average amount of information required for the class label of Tuples D. The larger the entropy is, the greater the uncertainty of the variable. If the training tuple D is divided according to the characteristic attribute A, the expected information of D is described as formula (7). (Note: The expected information of D is conditional entropy, which is based on the classification of characteristic attribute A).

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j) \tag{7}$$

Dj is the classification of feature attribute A, as well as v is the number of types of the characteristic attribute A.

b) Calculating the information gain.

The information gain is the difference between the two information entropy.

$$gain(A) = info(D) - info_A(D) \tag{8}$$

As the above formula (8) shows, gain(A) represents the amount of information obtained by classifying A as a node. The more information, the more important A is.

c) ID3 algorithm is used to establish each child node in the tree. According to the characteristics of the data set selected by the information gain, the algorithm selects the feature with the maximum information gain as the judgment node and acts as the sub-node in the tree.

d) Using recursive thinking, repeat above steps from (1) to (3) so as to establish the decision tree.

2) Bagging integrated decision tree

Bagging is a technology of repeated sampling from data according to uniform probability distribution. The algorithm does use the different training set to fit a single member classifier in the ensemble classifier, as well as Bootstrap sampling is used by training set in the fitting process. which is a random sampling with a rewind. So bagging can improve the accuracy of unstable model, and reduce the degree of over-fitting. The final result of the algorithm is to construct a series of prediction functions, and combining them into a prediction function by voting. The process is shown in figure 4.

The steps are described as follows.

a) The bootstrap [13] method is used to select n training samples from the sample set, and using it as the training set T1-Tn. This process is executing for K times, and k subsets $\{T_1, T_2 ... T_K\}$ are selected.

b) K sample subsets are trained on their own training data on all attributes. And then k classification models are obtained.

c) According to the classification model obtained by the above steps, the value of each {P1, P2 …, Pk} model is predicted respectively.

d) The value $\{P_1, P_2, \ldots, P_k\}$ of each model is combined by the average method. The final result is output. The formula of the averaging method is described as follows.

$$P(x) = \frac{1}{K} \sum_{i=1}^{K} p_i(x) \tag{9}$$

Hereby, pi is the value of a model. K is the number of samples training.

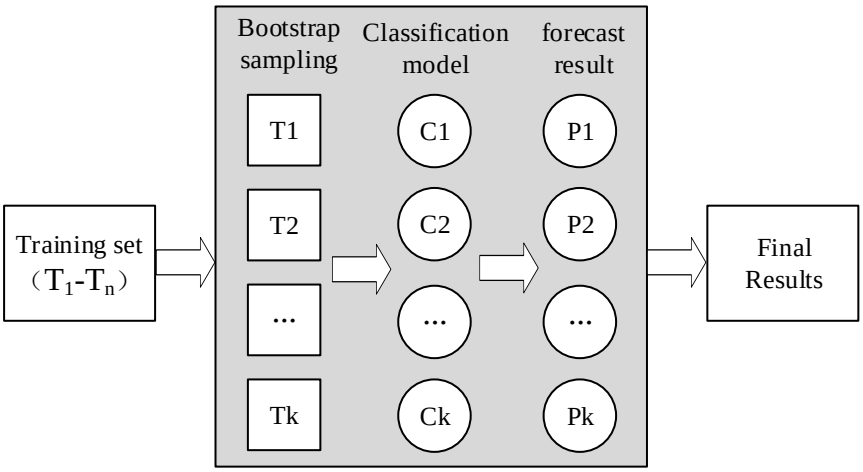


Figure 4. Bagging process

3) Gaussian Bayes algorithm

Compared with decision tree and Bagging algorithms, the greatest advantage is that when the large scale training set is selected, the Gauss Bayes algorithm only has relatively small number of features for each item, and the training and classification of the project is only a mathematical operation for the characteristic probability. Therefore, Gaussian Bayes algorithm has a fast speed when training large amount of data. The flow of this algorithm is shown as in Figure 5.

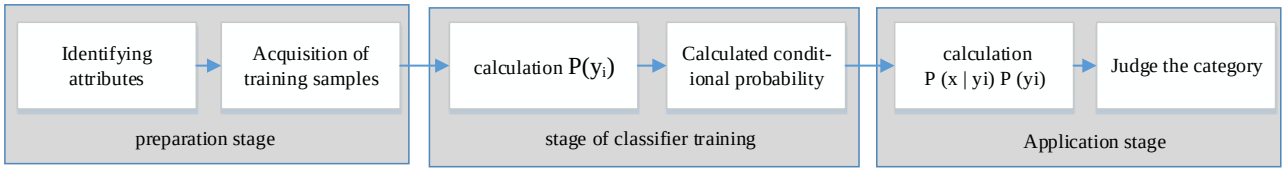


Figure 5. Gaussian Bayesian algorithm flow

As shown in figure 5, the entire algorithm flow can be divided into three phases.

The first is Preparation stage. This stage determines the characteristic attributes according to the specific situation, and dividing each feature attribute appropriately. And then, some of the items is classified by manually so as to form a training sample set. The input of this phase is all the data to be classified, and the output is the feature attribute and the training sample. This stage is the only stage that needs to be completed manually in the whole naive Bayesian classification. The quality of the classifier will have an important impact on the whole process, as well as be affected by the feature attributes, the classification of the feature attributes and the training samples.

The second is Classifier training stage. This stage is generating the classifier. Firstly, the occurrence frequency of each class is calculated in the training sample. And then the conditional probability of each category on the characteristic attribute is calculated. Finally, the results are recorded. The input is the feature attribute and the training sample, and the output is the classifier.

The third is Application stage. The task at this stage is to classify the classified items by using the classifier. The input of this stage is classifier and the item to be classified, and the output is the mapping relationship between items to be classified and categories. The specific steps are described as follows.

The specific steps are described as follows.

a) Assume $x = \{x_1, x_2, ..., x_m\}$ be an item to be classified, and each x_i be a characteristic attribute of x.

b) Setting have a set of y categories where $y_1=0, y_2=1$.

c) Calculating conditional probability $P(x_i \mid y_j)$ .The formula of (10) is shown as follows.

d) If it is existed as $P(x \mid y_k) = \max\{P(x_i \mid y_j)\}$, then it will be $x \in y_k$.

e) According to Bayesian theorem, the following formulas can be obtained. The formula of (11) is shown as follows.

f) The class that maximizes the value of P(x|yi) P(yi) is found out, and the items to be classified fall into this category.

$$P(x_i \mid y_j) = \{P(x_1 \mid y_1), P(x_2 \mid y_1), ..., P(x_m \mid y_1), P(x_1 \mid y_2), P(x_2 \mid y_2), ..., P(x_m \mid y_2)\} \quad (10)$$

$$P(x \mid y_i)P(y_i) = P(x_1 \mid y_j)P(x_2 \mid y_j)...P(x_m \mid y_j)P(y_j) = P(y_i)\prod_{i=1}^{m} P(x_i \mid y_j) \quad (11)$$

E. Building Model

The above three algorithms is adopted to construct the three model. Followed that, the principle of "the minority is subordinate to the majority " is used to reconstruct a new model. The model is called as intelligent reading model. The principle of "the minority is subordinate to the majority" means that, if a data is trained by three model, the output result of classification is "0 / 0 / 1". Because of the number of '0' in the result is more than the number of '1', the final result of the data is 0.

IV. VERIFICATION AND ANALYSIS OF THE MODEL

The data from the testing set are input into the intelligent reading model, as well as the processed results analyzed. The quality of the model is measured by two technical indicators, that is, Accuracy and F-Measure value.

The two-dimensional confusion matrix is shown in Table 2. The meaning of " the forecast is wrong, the actual is wrong (TN)" is that, the actual label category of the data is wrong, and it is still wrong after prediction. Based on the two-dimensional confusion matrix shown in Table 2, the formula (12) of the accuracy rate and the formula (13) of F-Measure are given as below.

TABLE II. TWO-DIMENSIONAL CONFUSION MATRIX

		actual value	
		positive	**Wrong**
Forecast value Such as type (10)	**positive**	The forecast is positive, The actual is positive (TP)	The forecast is positive, the actual is wrong (FP)
	Wrong	The forecast is wrong, the actual is positive (FN)	The forecast is wrong, the actual is wrong (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (12)$$

$$\text{F-Measure} = \frac{2*TP}{2*TP + FP + FN} \quad (13)$$

As shown in formula (12), the accuracy rate refers to the proportion of successful data in all predicted data. The predicted success means "the predicted value is same as the actual value". It includes two kinds of labels such as "the forecast is positive, the actual is positive (TP)" and "the forecast is wrong, the actual is wrong (TN)". When users ask some questions, they only want the right answers. Therefore, TN label is not necessary. As shown in formula (13), The F-Measure value is a comprehensive evaluation index of accuracy rate and recall rate. Because it does not include TN label, it is often used to evaluate the classification model.

Accuracy is a very objective evaluation index, but sometimes the accuracy rate does not represent the quality of the algorithm. Especially in the case of imbalance of positive and negative samples, the accuracy evaluation index has great defects. The most common F-Measure method is the weighted harmonic average of the accuracy rate and recall rate (the recall rate is the measure of the cover surface). Because the F-Measure method comprehensively considers the accuracy rate and recall rate, it effectively avoids the problem of unbalanced data distribution. Therefore, comparing with the accuracy rate, the f-measure method can better reflect the quality of the algorithm. Among them, the higher the value of F-Measure, indicating the better classification results of the corresponding algorithm. If we combine the three algorithms such as decision tree, Bagging and Gauss Bayes, the results of the combination algorithm and the single algorithm are shown in Table 3.

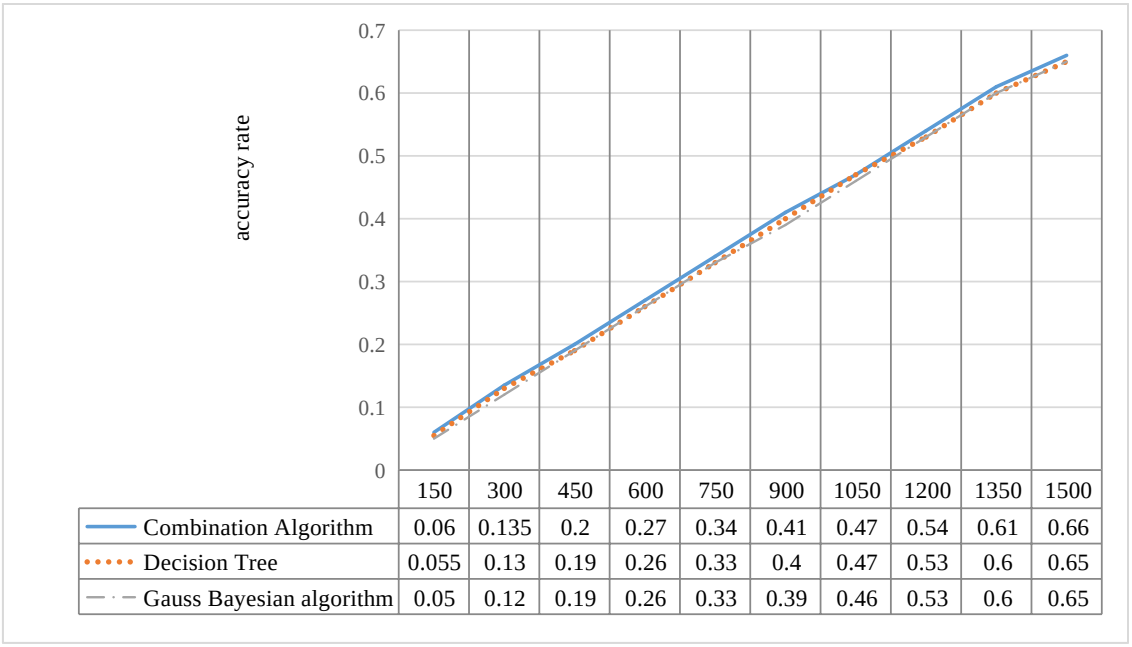
TABLE III. COMPARISON OF PREDICTION RESULTS

	decision tree	**Bagging**	**Gauss Bayesian**	**Combination algorithm**
Accuracy	0.6569	0.7105	0.7093	0.7112
F-Measure	0.3354	0.1933	0.2504	0.3381

Table 3 shows that the accuracy of the combination algorithm is 0.7112, while the other three separate algorithm accuracy is 0.6569, 0.6569 and 0.7105. Therefore, the combination algorithm is more accurate than the other three methods. In addition, the F-Measure value of the combined algorithm is 0.3381. The F-Measure value of the other three methods is 0.3354, 0.1933 and 0.2504. Therefore, the combination algorithm is better than the other three separate algorithms in terms of F-Measure.

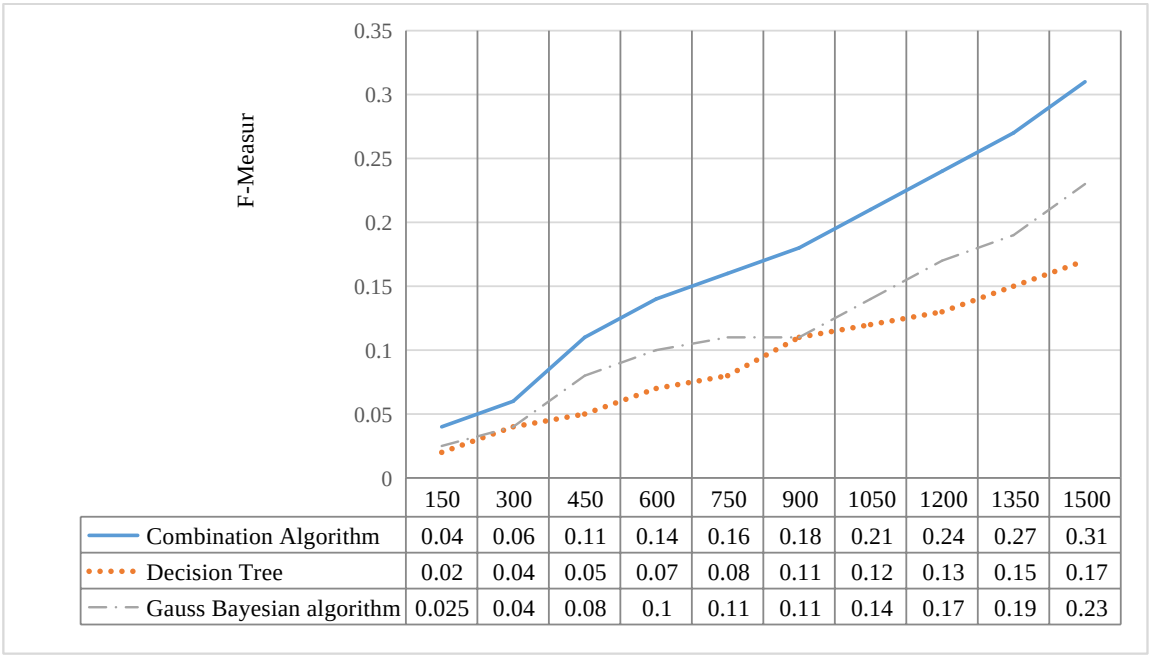
Whether it's accuracy or F-measure, the result of the combined algorithm is better than that of the other three separate algorithms. Therefore, the intelligent reading model is based on the combination of decision tree, Bagging and Gauss Bayesian algorithms.

In order to verify the superiority of the method, we are selecting about 8000 pieces of data for experimental verification. The experimental results are shown in figure 5 and figure 6.



	150	300	450	600	750	900	1050	1200	1350	1500
Combination Algorithm	0.06	0.135	0.2	0.27	0.34	0.41	0.47	0.54	0.61	0.66
Decision Tree	0.055	0.13	0.19	0.26	0.33	0.4	0.47	0.53	0.6	0.65
Gauss Bayesian algorithm	0.05	0.12	0.19	0.26	0.33	0.39	0.46	0.53	0.6	0.65

Figure 6. Accuracy comparison diagram



	150	300	450	600	750	900	1050	1200	1350	1500
Combination Algorithm	0.04	0.06	0.11	0.14	0.16	0.18	0.21	0.24	0.27	0.31
Decision Tree	0.02	0.04	0.05	0.07	0.08	0.11	0.12	0.13	0.15	0.17
Gauss Bayesian algorithm	0.025	0.04	0.08	0.1	0.11	0.11	0.14	0.17	0.19	0.23

Figure 7. F-Measure comparison

Experiment in figure 6, the X-axis represents the amount of data in the experiment, and the Y-axis represents accuracy, which shows that comparing with the three algorithms of decision tree, Bagging and Gauss Bayes, the combination algorithm has a slightly better accuracy. in figure 7, the X-axis represents the amount of data, and the Y-axis represents the F-Measure. The F-Measure of the combined algorithm is obviously superior to other three separate algorithms.

V. CONCLUSION

The reading model constructed in this paper makes reading more intelligent. Aiming at the problem of natural language input, the corresponding answer can be given according to the existing TXT content. According to the experimental data, it can be concluded that the intelligent reading model based on the combination of decision tree, Bagging and Gauss Bayesian algorithm has a good classification ability.

REFERENCES

[1] Xi Xuefeng, Zhou Guodong. A study of Deep Learning for Natural language processing [J]. Journal of Automation! 42 (10): 1445-146

[2] Chen Lian. Research and Application of key Interactive Technology based on Web Intelligent Education platform [D]. Graduate School of Chinese Academy of Sciences (Chengdu Institute of computer Application).

[3] Han Dongxu, Chang Baobao. Domain adaptability method of Chinese word Segmentation Model [J]. Journal of computer Science, China (02): 272-281.

[4] Yang Bin, Han Qingwen, Lei Min, Zhang Yapeng, Liu Xiangguo, Yang Yaqiang, Ma Xuefeng. Short text Classification algorithm based on improved TF-IDF weight [J]. Journal of Chongqing University of Technology (Natural Science) 30 (12): 108-113.

[5] Qian Cheng, Yang Xiaolan, Zhu Fuxi. Python-based web crawler technology [J]. Heilongjiang Science and Technology Information 2016 (36): 273.

[6] Wang Daoming, Lu Changhua, Jiang Weiwei, Xiao Mingxia, Li inevitable. Research on decision Tree SVM multiple Classification method based on Particle Swarm Optimization algorithm [J]. Journal of Electronic Measurement and Instruments 29 (04): 611-615.

[7] Bi Kai, Wang Xiaodan, Yao Xu, Zhou Jindeng. An adaptive selective integration based on bagging and confusion matrix [J]. Chinese Journal of Electronic Science (EJ). 42 (04): 711-716.

[8] Zhu Mingmin. Study on Bayesian Network Structure Learning and reasoning [D]. Xi'an University of Electronic Science and Technology.

[9] Zan Hongying, Zuo Weisong, Zhang Kunli, Wu Yunfang. A study on emotion Analysis combined with rules and Statistics [J]. Computer Engineering and Science 33 (05): 146-150.

[10] Gu Yijun, Fan Xiaozhong, Wang Jianhua, Wang Tao, Huang Weijin. Automatic selection of Chinese stops word list [J]. Beijing Institute of Technology Proceedings 2005 (04): 337-340.

[11] Liu Qinghe, Liang Zhengyou. A feature selection method based on information gain [J]. Computer Engineering and applications 47 (12): 130-132 + 136.

[12] Huang Yuda, Fan Taihua. Analysis and Optimization of decision Tree ID3 algorithm [J]. Computer Engineering and Design! 33 (08): 3089-3093.

[13] Liu Jian, Wu Yi, Tan Lu. Improvement of bootstrap method for self-help sampling [J]. Mathematical Theory and Application 2006 (01): 69-72.