www.cad-journal.net
Computer-Aided Design™
AND APPLICATIONS

Integrated Learning-based Multimodal Text Analysis Algorithm for College English Textbooks

Yan Ding[1], Wei Dong[2], Liang Lu[3] and Chunyi Lou[4]

[1]Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, dy@qvc.edu.cn
[2]Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, dw@qvc.edu.cn
[3]Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, ll@qvc.edu.cn
[4]Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, lcy@qvc.edu.cn

Corresponding author: Chunyi Lou, lcy@qvc.edu.cn

Abstract. With the deepening AI and big data are becoming more and more important to our production and life. We can carry out relevant operations through mobile terminals in our daily life. It can effectively improve the drawbacks of the traditional teaching model, fully mobilize students' learning interests, change the role, and effectively improve teaching quality and efficiency. The research is aimed at the effectiveness of computer assisted language. Through systematizing the process of text classification of English textbooks, this paper introduces the preprocessing, feature selection algorithm, similarity calculation, text representation and classifier algorithm of text classification of English textbooks. It also introduces several classifier models commonly used in the field of text classification in English textbooks: naive Bayes. The performance evaluation index of the classifier is introduced. In order to verify the effectiveness of the three optimization algorithms (OOB-WRF, Ada NB and Ada RCFNB) proposed in this paper in English text classification of English textbooks, this paper selects the English newsgroup corpus for experimental verification. F1 value of OOB-WRF algorithm are better than the traditional random forest algorithm.

Keywords: text classification of English textbooks; integrated learning; random forest; adaptive boosting; plain Bayesian; computer-aided and multimodal text analysis
DOI: https://doi.org/10.14733/cadaps.2023.S10.113-123

1 INTRODUCTION

Traditional college English teaching is mainly based on explanation and theory. There are not many teaching hours, which are supplemented by some listening and speaking training. It is very valuable to divide a part of the class hours to strengthen English practical teaching. Building a relatively independent English practical teaching system is a difficult problem to be solved in front of every educator. It is also an arduous and urgent task for higher vocational colleges to cultivate skilled talents. Vigorously strengthen the research - actively explore and build a vocational ability-oriented and talent training objectives. The English practical teaching system reflecting the characteristics. Combining computer-aided language in college English teaching has great convenience and can fully highlight the characteristics of information and technology. For example, when teaching English listening in colleges and universities, students are mainly provided with independent learning space at this stage. In the process of teaching, the type of computer-aided language teaching is also selected. Muaad et al. [1]. proposed a model recognition system based on Arabic text mechanism. Through different character spacing and text recognition, the understanding and representation of the text are split, and the character by character is analyzed. Finally, a practical solution is provided. For example, operational teaching allows students to practice the topics corresponding to the content of the textbook. In this way, you can find your own problems through long-term practice, so as to correct them in time. The relationship between teachers and students should also be appropriately changed. Teachers and students should talk to each other in class and discuss the listening content with students.

The most important thing to carry out English teaching is practicality, so as to prepare for entering the society in the future. In practice, the use of computer-aided language in English teaching can not only fully reflect the business theme of college English. It can also guide students to practice repeatedly and firmly grasp the English knowledge they have learned. Pan and Qin [2] computer-aided language can also provide a large number of materials for students' language learning. Improving learning enthusiasm Computer assisted language can make English teaching materials more intuitive and vivid, because students' interest in learning can be aroused through video and pictures. Especially for the interpretation teaching, after the teacher explains the English-Chinese interpretation skills, students can improve the learning effect through man-machine interpretation practice.

Gong [3] has established a new voice simulation teaching model with traditional corpus as innovation. Through the simulation of the big data language teaching model, the analysis and construction of the computer educational administration system are matched and analyzed. It automatically restores the management practice based on the teaching content. The performance evaluation index of the classifier is introduced. The theoretical basis of parallel integration method bagging and sequence integration method boosting in ensemble learning is systematically studied. The expansion algorithm of bagging "random forest" is introduced in detail. The simple voting method used in the result prediction makes all decision trees in the random forest have the same contribution to the final prediction result.

2 LITERATURE REVIEW

No matter what kind of course teaching is carried out, advanced technology and equipment such as multimedia courseware and computer software are required. However, it is teachers who occupy the core position, and the development of actual teaching still needs teachers' leading role. The most ideal state of computer-aided language environment is to cover the following three elements. After analyzing students' actual language expression, Shu [4] believes that students' awareness of activity freedom needs to be promoted by the external environment. In order to improve students' ability of learning activities, it is necessary to carry out good environmental strategy analysis for students. So as to develop good teaching awareness and learning ability, and improve the deficiencies in teaching methods. Among them, the application of multimedia computer has good advantages in the process of task-based language teaching. Through the incomplete analysis of application level, the problem of students' different level foundation is solved. Among the different models proposed by Muaad et al [5], the media application of each content of the model is inseparable. The description of multi-level characters requires means simulation under different programs. With the continuous development of computer in language application programming, it is possible to use the character structure under the design model for language education model. Jing and Jiang [6] introduced some very practical computer programming tools and carried out model domain analysis on a wide range of teaching media.

The curriculum reviews in the central classroom, thus enhancing the inference of learner satisfaction. Chen et al. [7] studied the differences of learners' complaints at the MOOC curriculum level. It indicates the main complaints of learners about high-level MOOC courses. Khafaga [8] has studied the effectiveness and application. The computer plays a very important role in the writing and expression of text, including the interpretation of script. It is very helpful for the preprocessing operations of text data in language learning, such as cleaning, word segmentation, and removing inactive words. In order to facilitate the subsequent analysis and processing, the computer can also divide the text into different categories or groups to find the internal laws and similarities of the text data. Hassanzadeh et al [9] classifies the text data by clustering the text, analyzes the emotional tendency and polarity in the text, so as to understand the user's emotional attitude and behavior. Themes are modeled for text data to find topics and topics in the text, so as to understand the internal meaning and background of the text. Extract the important information and core content in the text so that users can quickly understand the main content of the text [10]. Integrated learning mainly adopts the idea of "learning from the best", and predicts the same training data set by building multiple basic classifiers with good classification performance, which can improve the generalization performance and classification accuracy of classifiers. As one of the most popular machine learning methods, the theoretical validity of ensemble learning is still lacking. Therefore, designing an ensemble learning classifier with excellent performance and strong generalization ability is still a subject to be explored in the field of text classification of English textbooks.

3 APPLICATION AND OPTIMIZATION OF INTEGRATED LEARNING IN ENGLISH TEXT CLASSIFICATION OF ENGLISH TEXTBOOKS

The computer-aided language teaching system can effectively improve the convenience of teaching and improve the quality and efficiency of the overall teaching. However, when carrying out college English teaching, we must also make reasonable use of computer-assisted language teaching. Regularly organize teachers to learn relevant computer-aided language teaching technology. Colleges and universities should improve infrastructure construction, create a good the healthy and orderly development of colleges and universities. Ensemble learning is a very popular machine learning method that builds and combines multiple weak classifiers to accomplish a learning task. Compared with a single model, the base classifiers in ensemble learning provide better prediction results by using a "crowd-sourcing" approach. Due to the good generalization ability of integrated learning, it has been used in a wide range of applications, such as classification prediction, regression problems, feature selection, and outlier detection. The integrated learning approach is shown in Figure 1.

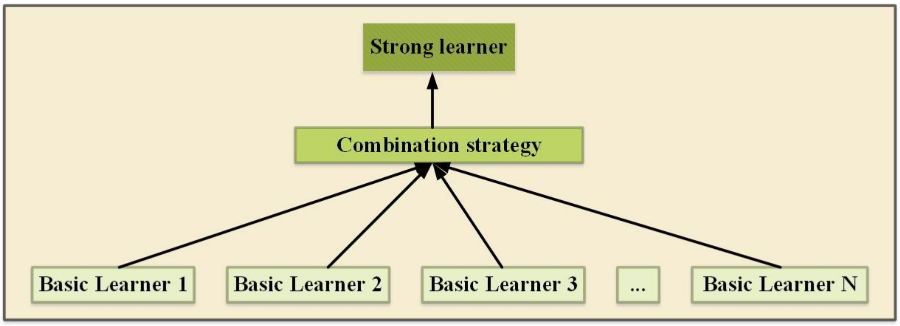

Figure 1: Schematic diagram of the integrated learning approach.

Bagging (Bootstrapped Aggregation) is an integrated by Beriman in 1996. This algorithm uses Bootstap sampling to obtain T self-sample datasets from a dataset D with M samples with put-back sampling. This sampling ensures that some samples in dataset D may appear multiple times while some samples do not appear in these T self-help sample datasets. The Bagging model is shown in Figure 2.

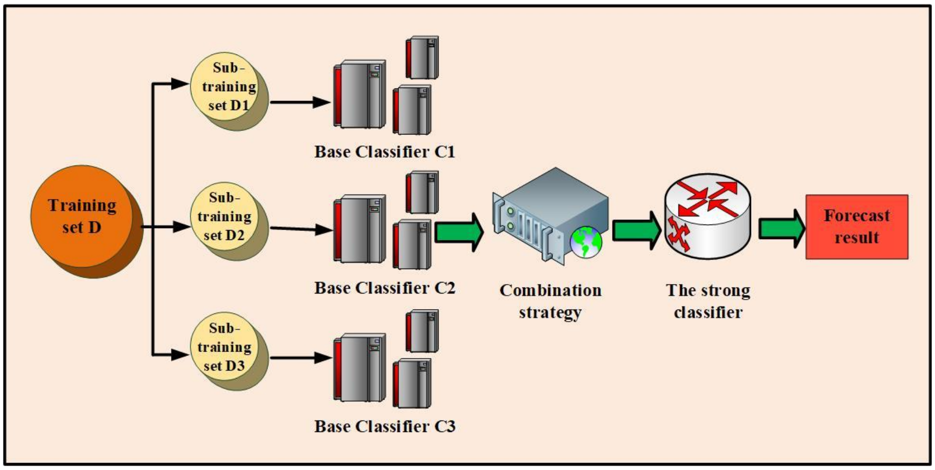

Figure 2: Bagging schematic.

Breiman points out that the performance of the Bagging algorithm is related to the stability of the base classifier, and we call the base classifier unstable if the training set changes very little but causes significant changes in the performance. At the same time, when the number of basic classifiers is T, the formula for calculating the weight of the ith classifier is given in this paper as.

$$\mathrm{weight\ (i)} = 1 - \frac{1/\mathrm{con}(i)}{\sum_{j=1}^{T} 1/\mathrm{con}(j)} - \frac{T-2}{T} \tag{3.1}$$

where, $\mathrm{con}(i)$ denotes the posterior probability of the i-th classifier.

For the above weight calculation formula, it is proved in the literature that the corresponding weight is larger when the classifier is more accurate.

$$\sum_{i=1}^{T} weight\left(i\right) = 1 \tag{3.2}$$

The random forest algorithm as a classical algorithm in bagging, is here optimized by using the above theory, to give weights to each decision tree in the random forest, but because the posterior probability con(i) calculated according to the Bayesian formula is more complicated, here we use oob error. For the training process in the random forest for the training set with put-back, Breiman pointed out that the use of out-of-bag samples is very effective for the optimization of the algorithm when he pointed out the bagging algorithm. The calculation formula is.

$$\mathrm{oob_{e}rr(i)=err_{n}um(i)} \tag{3.3}$$

where, e^{2rr} num(i) denotes the number of classification errors of the i-th base classifier for out-of-bag samples. Obb_num(i) is the number of out-of-bag samples.

The generalization error rate estimated by out-of-bag oob_err is the error rate on the training set. That is, it is an unbiased estimate of the base classifier. In this case, oob_err(i) can be used instead of the above posterior probability con(i) to simplify the calculation.

$$obb\_\,\mathrm{err(i)} \approx 1 - \mathrm{con}(i) \tag{3.4}$$

The above weight formula can be deformed as:

$$\mathrm{weight(i)} = 1 - \frac{1/\left[1-\mathrm{obb}_{\mathrm{err(i)}}\right]}{\left.\sum_{j=1}^{T}\frac{1}{\left[1-\mathrm{obberrr}_{\mathrm{er}}\right]}\right]} - \frac{T-2}{T} = \frac{2}{T} - \frac{1/\left[1-\mathrm{obb}_{\mathrm{err(i}}\right]}{\sum_{j=1}^{T} 1/\left[1-\mathrm{obbb}_{\mathrm{err(j)}}\right]} \tag{3.5}$$

AdaBoost (Adaptive Boosting) is one of the most popular Boosting algorithms. This algorithm can continuously add new weak classifiers during the training process according to the demand of accuracy, and the weight of the training sample indicates the possibility of being selected for the next round of training.

Step1: Input a training set of size n.

$$\mathrm{T} = \{(\mathrm{t1},\mathrm{y1}),(\mathrm{t2},\mathrm{y2}),\ldots,(\mathrm{tn},\mathrm{yn})\} \tag{3.6}$$

where y is the category label of the corresponding instance t.

Step2:

$$\mathrm{w1} = (\mathrm{w11},\mathrm{w12},\ldots,\mathrm{w1i},\ldots,\mathrm{w1n}) \tag{3.7}$$

$$\mathrm{w1i} = 1/\mathrm{n}_{\text{¡£}} \tag{3.8}$$

Step3: Randomly select p samples with put-back from T based on the weight vector w1. The weight matrix is.

$$w_h = (\mathrm{wh1},\mathrm{wh2},\ldots,\mathrm{w1i},\ldots,\mathrm{whp}) \tag{3.9}$$

Step4: calculate the weighted error function e_h during the training process :

$$e_h = P\left(h\left(t_i\right) \neq y_i\right) = \sum_{i=1}^{n} w_{hi} I\left(\neq y_i\right) \tag{3.10}$$

Step5:

$$\alpha_h = \frac{1}{2}\ln\frac{1-e_h}{e_h}\ldots \tag{3.11}$$

Step6:

$$w_{h+1} = \left(w_{h+1,1}, w_{h+1,2}, \ldots, w_{h+1,i}, \ldots, w_{h+1,n}\right) \tag{3.12}$$

$$w_{h+1,i} = \frac{w_{h,i}}{z_m}\exp\left(-\alpha_h y_i h\left(t_i\right)\right) \quad , \quad i = 1,2,\ldots,n \tag{3.13}$$

Where Z_m means:

$$Z_m = \sum_{i=1}^{n} w_{h,i}\exp\left(-\alpha_h y_i h\left(t_i\right)\right) \tag{3.14}$$

Step7: Repeat steps 3-6 above.

Step8:

$$\mathrm{H(t)} = \mathrm{sign}\left(\sum_{k=1}^{K}\alpha_k h_k(t)\right) \tag{3.15}$$

In most cases, the AdaBoost algorithm performs well enough.

4 ANALYSIS OF SIMULATION RESULTS

Computer-aided speech teaching is an integral part of computer-assisted language learning. With the emergence of modern technology, voice teaching has undergone major changes and eliminated some limitations. English spelling teachers can provide learners with a rich learning environment. The computer-aided voice teaching program is helpful to cultivate personalized, diverse and comfortable virtual learning environment. There are many advantages in ensuring the learnability of the target language voice system and improving learners' pronunciation skills. Learners can access endless voice input and personalized feedback. The high-quality image effect of computer-assisted speech learning software allows learners to see the pronunciation actions in the process of pronunciation, and allows learners to compare the pronunciation of native speakers with their own. Computer-aided speech learning software provides a zero-pressure environment, allowing learners to actively participate in unlimited speech input at their own pace. It also receives feedback on individual learners' speech learning to help learners train rhythm and improve their speech level. Four sets of comparison experiments are designed for the effectiveness of the integrated learning algorithms introduced in Chapter 3 of this paper, including bagging, Adaboost and the corresponding optimization algorithms OOB-WRF, Ada-NB and Ada-RCFNB. The comparison experiments are as follows: (1) Comparison experiments of commonly used text classification of English textbooks algorithms. (2) Decision tree, random forest, and OOB-WRF comparison experiments. (3) Parsimonious Bayes, Ada-NB, and Ada-RCFNB.

In this paper, the Newsgroups corpus is chosen for the experimental corpus. These documents are evenly distributed among 20 newsgroups, also known as the 20newsgroups dataset. There are currently three versions of this dataset: (1) the original dataset: 20news-19997.tar.gz. (2) the dataset divided into two parts in chronological order for training and testing: 20news-bydate.tar.gz. (3) the one processed by Jason Rennie at MIT University, USA, which does not contain duplicate documents, but only single source and topic. Each document belongs to only one newsgroup: 20news-18828.tar.gz. Its topics contain 6 categories (electronics, sports, science, music, politics, religion) and 20 newsgroups in sklearn, and this dataset is loaded by sklearn. datasets. fetch_20newsgroups. There are 18846 texts in the dataset, of which 11314 are training texts and 7532 are testing texts, divided according to 6:4.

This experiment compares three commonly used text classification of English textbooks algorithms including: KNN, Bayesian classifier, SVM, and verifies the Precision and Recall and F1 values of documents before and after pre-processing by comparing all topics (20) of a 20-group news corpus.

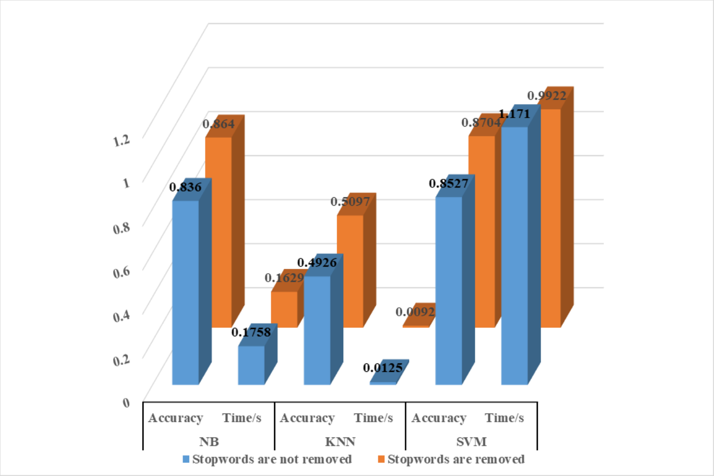

Figure 3: Comparison of classifier performance before and after preprocessing.

From Figure 3, it can be seen that the accuracy of these three classifiers SVM>NB>KNN, while the computing time SVM>>NB>KNN. It can be seen that although the SVM model has a high degree of classification accuracy, the computing time is also very large. Comparing before and after removing the deactivated words, the accuracy is improved, which shows the necessity of the pre-processing stage in the text classification of English textbooks process, not only to improve the model classification accuracy, but also to improve the computing speed. From the above introduction, we can see that this dataset contains 20 categories, but some of the categories are not very different from each other, and it is inconvenient to observe too many categories, so we selected one category in each category for classification: rec.autos, talk.religion.misc, and 6 categories. Here we use the Tf-idf method for feature extraction, and use the NB classifier with good accuracy and fast operation speed in the above experiments to classify the dataset, and get the results as shown in Figure 4.

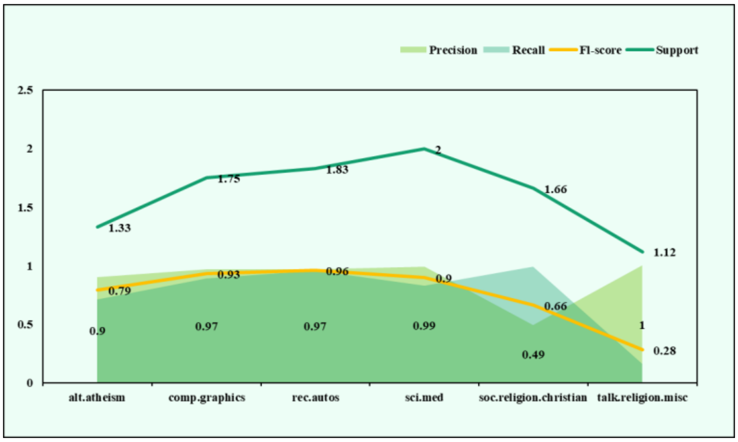

Figure 4: NB_TFIDF classification result matrix.

In this experiment, six types of text in the dataset are classified with different number of decision trees from 50 to 500, and random sampling with put-back is used in the training process, and out-of-bag is used as the test sample to compare and calculate the classification accuracy of C4.5 decision tree, RF and OOB_RF algorithms. As shown in Figure 5.

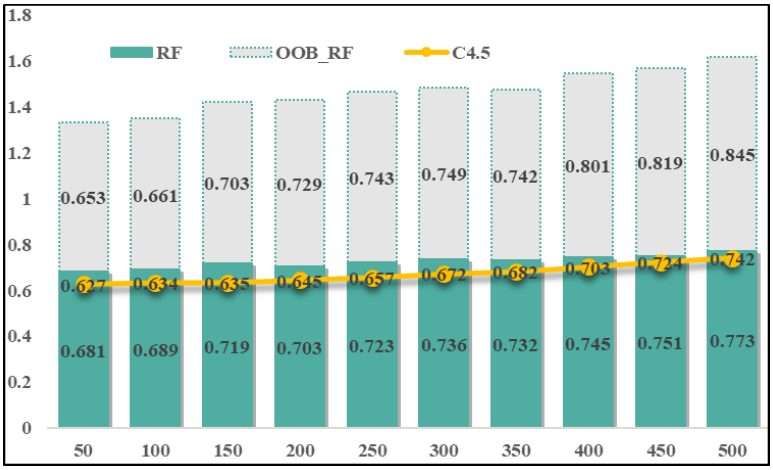

Figure 5: Accuracy of C4.5, RF, OOB_WRF under different decision trees.

The classification effect is unstable and oscillating; the classification accuracy of C4.5 decision tree algorithm with simple minority following majority is low, while our proposed OOB_WRF algorithm gradually increases with the increase of the number of base decision trees. The final classification accuracy can reach more than 85%.

When the number of decision trees is 100 - 500, the other three performance evaluation methods are combined here, as shown in Figure 6.

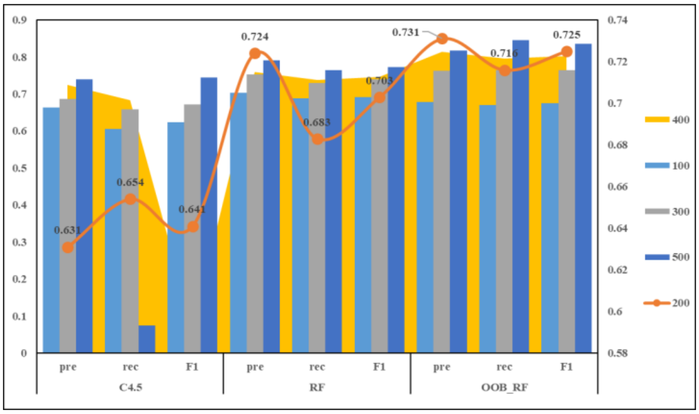

Figure 6: C4.5, RF, OOB_WRF performance comparison.

As seen in Figure 6 C4.5, RF, and OOB_WRF, it can be seen from the accuracy, recall, and the F1-score of the combined two: the classification gets better as the number of decision trees increases. Classifier performance OOB_RF>RF>C4.5.

In this paper, a total of 969 texts are classified in six categories, where support represents the number of texts in each category. The confusion matrix has 6 rows and 6 columns, each row representing a category; for example, the 6 numbers in the first row represent the number of 133 texts in the alt. atheism category classified into each of the 6 categories, among which 94 texts are correctly classified and 39 texts are incorrectly classified into the soc. religion. christian category. The diagonal elements of the matrix indicate the number of texts correctly classified in each category. The same preprocessing operation is carried out for the RCFNB classifier proposed in this paper in the following, and the RCFNB parameters are set in the experiment: the number of iterations is 3, weight=1, and the results are shown in Figure 7.

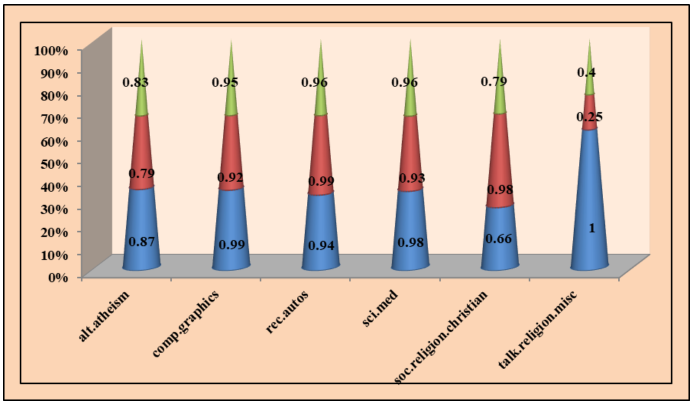

Figure 7: RCFNB classification results.

By optimizing the Bayesian classifier, the RCFNB classification accuracy is improved to 86.068%, and the recall as well as f1-score are also improved. From the confusion matrix, it is observed that the number of correctly classified texts for each category of the NB algorithm is 94, 156, 176, 166, 165, 18, while the number of correctly classified texts for each category of the RCFNB algorithm is 104, 172, 168, 171, 194, 25. The increase in the number of correctly classified texts can be observed very intuitively.

Finally, the same preprocessing operation is performed for the adaptive boosting algorithm Ada-RCFNB of RCFNB, and the classification results are shown in Figure 8.

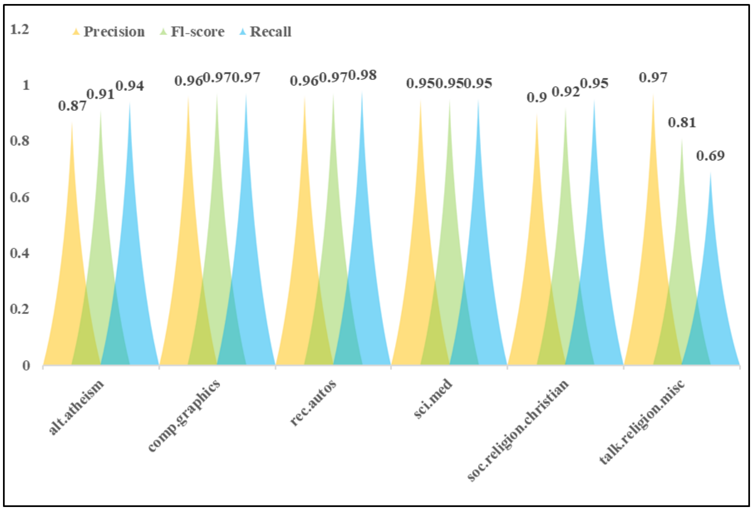

Figure 8: Ada-RCFNB classification results.

After adaptive boosting optimization of the RCFNB algorithm, the classification accuracy of the Ada-RCFNB classifier can reach 93.2%. The confusion matrix shows that for talk.religion.misc, which is the least correctly classified category, there are 25 correctly classified articles in the RCFNB algorithm, while the adaptive boosting algorithm can reach 69 articles. Therefore, the Ada-RCFNB algorithm is effective for the optimization of the RCFNB algorithm. Therefore. For NB, RCFNB, and Ada-RCFNB, the performance of RCFNB in terms of accuracy, recall, and F1 value is improved compared to the NB classifier, indicating that RCFNB is effectively optimized for the traditional Bayesian algorithm. And Ada-RCFNB further optimizes the RCFNB algorithm using the idea of adaptive boosting. In the following, the NB, RCFNB, Ada-NB, Ada-RCFNB, KNN and SVM classifiers proposed in this paper are compared and tested. Six classes of texts with different labels are selected, and the experiments are performed by cross-validation method, using f1-score as the performance evaluation index to obtain the classifier contrast, as shown in Figure 9.

Figure 9 gives the change curves of f1-score values of NB, RCF-NB, Ada-NB, Ada-RCFNB, KNN and SVM. It can be seen that the classification effect of Ada-NB and RCF-NB is better than that of NB, and the integrated learning plays a certain role in optimizing the base classifier, and the f1-score of adaptive boosting algorithm Ada-RCFNB is better than that of RCF-NB. In most cases, the Ada-RCFNB algorithm classifies better than SVM.

5 SUMMARY AND OUTLOOK

The in-depth study of ensemble learning is selected, which is usually used for classification prediction, regression problems, feature selection and outlier detection to improve the algorithm performance.

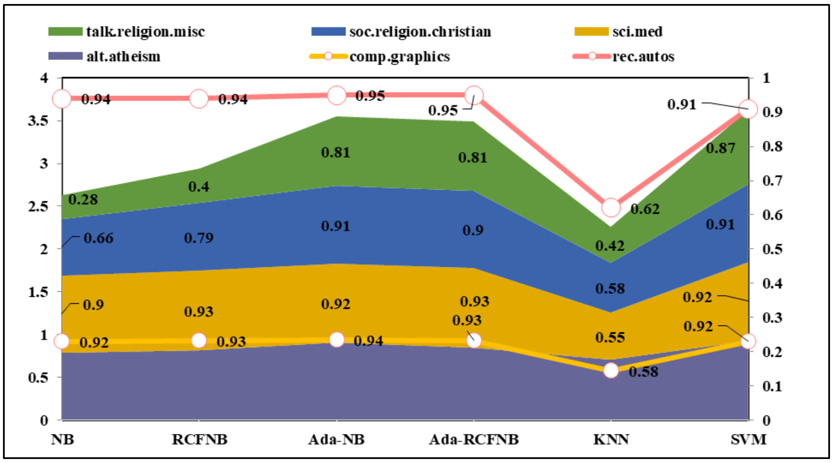

Figure 9: The classifier contrast.

In addition, this paper also studies the integrated learning algorithm applied to English text classification of English textbooks, and proposes an improved algorithm of the basic integrated learning model. Although it shows good results in classification accuracy, the increase of data volume has put forward requirements for the performance of the above classification algorithms and the popularization of the scope of application: (1) The integration optimization of Bayesian classifier in this paper improves the algorithm (1) While improving the accuracy of the algorithm, the integration optimization of Bayesian classification in this paper sacrifices time. (2) The integrated learning algorithm of training sets with multiple rounds of training may lead to over-fitting. How to select the number of classifier training rounds through theoretical analysis. (3) The integrated learning and optimization algorithms proposed in this paper are aimed at single-label English data sets, and extend the application of the algorithm in Chinese text and multi-label text.

Yan Ding, https://orcid.org/0000-0001-5170-628X
Wei Dong, https://orcid.org/0000-0002-9355-6811
Liang Lu, https://orcid.org/0009-0004-9492-6038
Chunyi Lou, https://orcid.org/0000-0003-3167-2021

REFERENCES

[1] Muaad, A. Y.; Jayappa, H.; Al-Antari, M.-A.; Lee, S.: ArCAR: a novel deep learning computer-aided recognition for character-level Arabic text representation and recognition, Algorithms, 14(7), 2021, 216. https://doi.org/10.3390/a14070216

[2] Pan, B.; Qin, Q.: Construction of parallel corpus for english translation teaching based on computer aided translation software, Computer-Aided Design and Applications, 19(1), 2022, 70-80. http://doi.org/10.14733/cadaps.2022.S1.70-80

[3] Gong, W.: An innovative English teaching system based on computer aided technology and corpus management, International Journal of Emerging Technologies in Learning (Online), 14(14), 2019, 69. http://doi.org/10.3991/ijet.v14i14.10817

[4] Shu, Y.: Experimental data analysis of college English teaching based on computer multimedia technology, Computer-Aided Design and Applications, 17(S2), 2020, 46-56. http://doi.org/10.14733/cadaps.2020.S2.46-56

[5] Muaad, A.-Y.; Kumar, G.-H.; Hanumanthappa, J.; Benifa, J.-B.; Mourya, M.-N.; Chola, C.; Bhairava, R.: An effective approach for Arabic document classification using machine learning,

Global Transitions Proceedings, 3(1), 2022, 267-271. https://doi.org/10.1016/j.gltp.2022.03.003

[6] Jing, D.; Jiang, X.: Optimization of computer-aided English teaching system realized by VB software, Computer-Aided Design and Applications, 19(S1), 2021, 139-150. https://doi.org/10.14733/cadaps.2022.S1.139-150

[7] Chen, X.; Cheng, G.; Xie, H.; Chen, G.; Zou, D.: Understanding MOOC reviews: Text mining using structural topic model, Human-Centric Intelligent Systems, 1(3-4), 2021, 55-65. https://doi.org/10.2991/hcis.k.211118.001

[8] Khafaga, A.-F.: The Effectiveness of CATA Software in Exploring the Significance of Modal Verbs in Large Data Texts, International Journal of Advanced Computer Science and Applications, 13(2), 2022, 1. https://doi.org/10.14569/IJACSA.2022.0130292

[9] Hassanzadeh, M.; Saffari, E.; Rezaei, S.: The impact of computer-aided concept mapping on EFL learners' lexical diversity: A process writing experiment, ReCALL, 33(3), 2021, 214-228. https://doi.org/10.1017/S095834402100001X

[10] Gao, Y.: Computer-aided instruction in college English teaching under the network environment, Computer-Aided Design and Applications, 18(S4), 2021, 141-151. https://doi.org/10.14733/cadaps. 2021 s4.141-151