Research Article

Construction of Intelligent Evaluation Model of English Composition Based on Machine Learning

Yuling Wang

Nanchong Vocational and Technical College, Nanchong 637131, China

Correspondence should be addressed to Yuling Wang; 1931011309@siit.edu.cn

Received 23 April 2022; Revised 4 June 2022; Accepted 6 June 2022; Published 8 July 2022

Academic Editor: Liping Zhang

Copyright © 2022 Yuling Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To automatically grade compositions, automatic composition grading primarily employs statistics, mathematical analysis, machine learning, natural language processing, and other technologies. This paper presents a machine learning-based intelligent evaluation model for English composition. As a result, this paper proposes a word vector grouping-based text content representation method and a vector space model-based text content representation method. The Word2Vec model is first trained with words, then used to generate a test model of word vectors, with the statistical information of the corresponding words in each category serving as content text features. The results show that when features based on word vector clustering are added to the content text, the effect of each model improves significantly, especially the maximum entropy model, which improves by 0.048. The XGBoost model has also seen a significant improvement, going from 0.771 to 0.803. The test corpus is made up of 100 articles chosen at random from the corpus, and the test set is checked for errors. The rate of accuracy is 68 percent. The conclusion demonstrates that the model presented in this paper can help promote English teaching reform and quality-oriented higher education while also easing the burden on teachers and students.

1. Introduction

English composition can reflect students’ writing, thinking, and analytical abilities, and it is a necessary and important examination content for standard examination training and various English exams. The traditional paper exam cannot keep up with the demands of today’s exams. This traditional examination not only takes a long time, but it also wastes a lot of paper and costs a lot of money. Artificial intelligence technology is having a significant impact on educational concepts, teaching models, and examination methods in the field of education. Intelligent correction is compatible with a wide range of composition scenarios, including online submission, scanning and recognition of handwritten answer sheets, as well as testing, homework, and other modes. Intelligent correction will convert student compositions into electronic format right away. The expert system based on knowledge base allows computers to do many things that intelligent teachers can do, and the speed with which they can do them has increased countless times and the scope has expanded countless times [1].

Front-line teachers are frequently troubled by the problem of correcting English compositions because they are overworked and lack the time and energy to do so. According to Yuan, we must first solve the problem of composition evaluation and correction before we can solve the problem of English writing teaching [2]. Liu and Li used quantitative linguistic features extracted from compositions as indicators of composition quality [3]. However, because the system has only one evaluation angle, it is entirely reliant on the statistical results of scoring indicators provided by experts and does not directly assess the composition’s internal quality, resulting in skewed scoring results. Smali et al. used Text Rank to prerank article quality before using a classifier to grade them [4]. This method can successfully identify candidates’ use of high-quality phrases, but it cannot extract deep semantic features. To extract text features for text clustering, Ban and Ning used Word2Vec word vector and convolutional neural network [5]. Liu et al. proposed a method for calculating the correlation score between the composition being tested and the topic based on rich features. Because this method requires training for various
composition topics in order to obtain various features, it has some practical limitations [6]. Random indexing is a technique proposed by Ondrejs and Gurcik. By incrementally accumulating text vectors in final gathering, random indexing obtains the corpus’s co-occurrence matrix, which takes very little time to construct. Meanwhile, using the incremental method, accurate word vectors can be obtained using only a small amount of text information [7].

Automatic computer grading can quickly obtain grading results and related feedback, allowing students to receive timely feedback and saving valuable time. Because different students require different learning guidance and learning goals when using software, it is necessary to learn software in order to make individualized diagnoses of students’ reactions and provide targeted learning guidance for different students. Because current automatic scoring methods have flaws, we use machine learning to assess the quality of compositions using semantic information from articles and machine learning [8, 9]. In comparison to the existing composition scoring system, it is equivalent to scoring the composition’s main ideas and central ideas expressed in the composition, which is closer to the scoring method used by real raters.

The research contribution of this paper is as follows:

(1) The article and preposition checking module based on the ML algorithm is designed, and the grammar checking problem is equivalent to the classification problem. After comparing several common classifiers, it is determined that the maximum entropy model is used to establish the prediction model and provide the function of error correction and judgment.

(2) Making an automatic grading system for English compositions. And selected different ML models to determine the types of compositions for off-topic analysis, as well as designing and extracting a large number of nontext features that can reflect the quality of students’ compositions from the side. On the basis of English composition representation methods, algorithms for detecting different types of off-topic compositions and analyzing the degree of off-topic compositions were developed.

2. Related Work

2.1. A Probe into Intelligent Evaluation of English Composition. In an English test, the score of composition consumes the most energy. In the process of evaluation, we should pay attention to the correctness of words, grammatical structure, and context in the composition. For grading teachers, not only should grading be reasonable, but it is also very time-consuming. Some scholar pointed out that the automatic scoring system uses approximate variables and intrinsic variables to get scores. Atkinson et al. put forward an LSTM (Long Short-Term Memory) to grade English compositions [10]. One of its remarkable advantages is that the whole process does not require developers to manually extract features, and there is no feature engineering in the whole process. The results show that it has achieved good results. Bandhakavi et al. developed an automatic composition scoring system based on latent semantic analysis [11]. Bozanis and Houmetis used a large number of topic models generated by training non-composition corpora to predict the topic distribution of the composition to be tested, extracted the corresponding words from the topic distribution as the topic words of the composition to be tested, and on this basis, conducted off-topic analysis through the matching degree of the topic words between the composition to be tested and the model essay [12].

One of the most important problems in the correction of English compositions is the grammatical problems in the compositions. English grammar is flexible, difficult to master, the level of college English is uneven, and the quality of composition is uneven, all of which bring difficulties to the examination of English grammar. Lo and Lo developed the first composition scoring system, which extracted the surface features of the text, such as word length and pronoun number, and used multiple linear regression to predict the composition score [13]. Narudin et al. evaluated the scores of compositions according to their semantic information. For each essay to be graded, the implicit semantic features of the essay are obtained through implicit semantic analysis, and the essay score is predicted by calculating the similarity with the semantic feature vectors of the already graded essays [14]. Chen and Asch proposed a hierarchical clustering method based on n-gram language model [15]. This method means that we can select multiple categories in the word category hierarchy to represent word categories, which can balance the situation that the expressive force of a certain category is insufficient due to the small number of words.

2.2. ML Technology Research. The ML algorithm has great practical value in academia and industry. Due to the quantity and complexity of big data, many traditional ML algorithms for small data are no longer suitable for the application of big data. Therefore, the research of ML algorithm in big data environment has become a topic of common concern in academia and industry.

Brockherde et al. put forward some statistical reasoning methods about big data [16]. When the divide-and-conquer algorithm is used to deal with statistical reasoning problems, it is necessary to obtain confidence intervals from huge data sets. Zhou et al. solved how to accurately predict online feature selection by using a small and fixed number of active features by studying sparse regularization and truncation techniques [17]. Fontana et al. compared some existing feature selection methods, proposed a distributed feature label model. In this model, the number of available features is the same as that in real data. Their test results on some high-dimensional data show that different feature selection algorithms have different performance under different model conditions, and they are related to the number of samples, global and heterogeneous tags [18].

Mohr et al. proposed a feature selection algorithm for classification [19]. The algorithm uses the local learning
theory to transform the complex nonlinear problem into a set of linear problems at first and then learns the feature relevance in the framework of maximum interval. Heung et al. use English web pages to mark information to solve the problem of cross-language classification, and put forward a method based on information bottleneck [20]. This method first translates Chinese into English and then encodes all web pages through an information bottleneck that only allows limited information to pass through. This method can make cross-language classification more accurate and significantly improve the accuracy of some existing supervised and semisupervised classifiers.

3. Methodology

3.1. Overall Design of English Composition Intelligent Evaluation Model. In the traditional question lottery, each student is randomly selected from each question in the question library. The question type must have a difficulty value. For students, the questions that can be asked are simple. On the other hand, this may be a difficult problem, so although this method of topic selection has greatly prevented cheating, it still affects the fairness of the examination. The intelligent evaluation model for English composition can generate personalized individual and class grading reports in real time and display them in the teacher's window. Intelligent grading is used for composition evaluation and data analysis in exams, which relieves teachers of much of their workload, not only by reducing the burden of correcting teachers' compositions, but also by allowing teachers to provide more personalized guidance to students. With the final composition score, each dimension has linear correlation, monotonicity, independence, inclusiveness, and balance. The composition to be graded is then graded according to each dimension, yielding multiple grading results. The corpus contains a database of sentence paragraphs and language points from all articles, which can be updated in real time. When the final grade results are announced, the relevant feedback (including sentences, paragraphs, and general feedback) is updated in real time, and learners can use this feedback to improve their language skills.

In natural language, word is the most basic unit, which has complete linguistic meaning and can be used independently. Therefore, word processing is fundamental and important in natural language processing. All words in natural language are classified according to their specific meanings and grammatical features, and the results are called parts of speech, such as nouns, verbs, adverbs, adjectives, and prepositions. English word segmentation is much simpler than Chinese, because English words are often separated by spaces. Use the collection of these words as the basis of sentence research. Therefore, as the main link of natural language research, whether a sentence can be reasonably and correctly decomposed into words will affect the overall effect of the system.

College English composition, in terms of testing, focuses on students' overall ability to use English language knowledge, such as spelling, word placement, grammar, word selection, and sentence construction, as well as understanding of essentials, design planning, and rhetorical style. Because of users' differing thinking habits and the ambiguity and flexibility of language, computers must understand natural language, which necessitates the use of complex rules in natural language processing technology, dynamic programming algorithms, and so on. The author tries to combine these two methods in order to provide effective composition evaluation results based on the characteristics of college English composition writing. The basic flowchart of English composition intelligent evaluation model is shown in Figure 1.

The basic diagnostic steps can be defined as follows:

1. Build a corpus. The author will build a grammar and vocabulary corpus for college courses, and a rule base for the structure of key sentences and keywords in the composition.
2. Unify all the rules of each key sentence and its similar sentences in each composition.
3. Filter out sentences with big differences by text similarity.
4. Match the regularly filtered sentences to the corresponding rule base, find the minimum editing distance, judge whether the wrong words are deformable, etc., and get the rule set with the highest total score.
5. Record the knowledge points corresponding to the rule with the highest total score, and repeat the processing of the next sentence until all the rules are exhausted.

3.2. Automatic Grading of English Compositions. Progress has been made in part-of-speech tagging, parsing, and word representation thanks to the ongoing development of natural language processing technology. There have been proposed automatic composition scoring methods based on statistics and natural language processing. Automatic composition grading, unlike manual grading, does not actually understand the composition, but rather evaluates it indirectly by constructing features that reflect the quality of words, sentences, and topics. To achieve this goal, the first step is to determine which elements reflect students' writing level, such as the number of common spelling mistakes, clause proficiency, and writing relevance to the topic. Second, once the writing scoring standard has been established, how to accurately and automatically extract relevant information from students' compositions by computer is dependent not only on related English ontology research, but also on the level of development of natural language processing technology. One difficulty of the traditional language model, when compared to other learning problems, is dimension disaster, which becomes more apparent when the joint probability of multiple independent variables is modeled. Word vectors trained with Word2Vec can not only express grammatical and semantic information about words, reveal language rules, and drastically reduce model training time.
Word2Vec has two training models: CBOW (Continues Bag-of-Words) model and Skip-Gram model. In this model, each word corresponds to a unique word vector. Given the word sequence \( D = \{w_1, w_2, \ldots, w_M\} \), the goal of the model is to maximize the average log probability:

\[
I(D) = \frac{1}{M} \sum_{i=1}^{M} \sum_{-k \leq c \leq k, c \neq 0} \log \Pr(w_i | w_c),
\]

where \( k \) is the sliding window size.

Word2Vec and GloVe vocabularies have both overlapping and nonoverlapping parts. The first step of our merger is to unify the vocabularies of the two, that is, to address the nonoverlapping parts. We express the word vector of this vocabulary as \( v_m \) in GloVe semantic space and define the calculation formula of vector \( v_m \) as follows:

\[
v_m = \frac{\sum_{i \in \text{nei} \times \text{Sim}(w_i, e_i)} v_i}{n},
\]

where \( \text{Sim}(w_i, e_i) \) is the cosine similarity between the word vector \( v_i \) and the word vector \( e_i \). Through the above formula, we can get the word vector representation of words that only exist in Word2Vec in GloVe semantic space, and vice versa.

In the English writing test, it is easier to get high marks by using synonyms with key words to express topics in various ways, instead of using the same key words to express topics throughout the text. Therefore, when constructing the mixed semantic space model of expression combination, this paper uses thesaurus to improve the distributed semantic space. Schematic diagram of the improvement process is shown in Figure 2.

The improvement process can be described as follows:

1. If two words belong to the synonym pair \((w_i, w_j)\) in the common synonym set in the composition, the Euclidean distance between the corresponding vectors of these two words in the distributed semantic space will be shortened.
(2) If two words belong to the synonym pair \((w_i, w_j)\) in WordNet synonym set, the distance between the corresponding vectors of these two words in distributed semantic space will be narrowed.

(3) Keep the to-be-speculated vector \(q_i\) of the word \(w_i\) and its initial vector \(w_i\) in the distributed semantic space.

The objective can be achieved by minimizing the objective function in the following formula:

\[
\psi(Q) = \sum_{i=1}^{l} \sum_{(w_i, w_{i+1}) \in E_i} \alpha_i \|q_i - \tilde{q}_i\|^2 + \sum_{(w_i, w_{i+1}) \in E_i} \beta_{ij} \|q_i - \tilde{q}_{ij}\|^2
\]

\[
+ \sum_{(w_i, w_{i+1}) \in E_i} \delta_{ij} \|q_i - \tilde{q}_{ij}\|^2
\]

where we define the distance between vectors as Euclidean distance, and \(\alpha_i, \beta_{ij}, \delta_{ij}\) as hyperparameter.

The maximum entropy model is a statistical learning model derived from the maximum entropy principle. Maximum entropy principle is the learning criterion of the learning probability model. When we apply it to classification problems, we get the maximum entropy model. It is assumed that the set of conditional probability models satisfying all constraints is

\[
C = \{P \in P | E_P(f_i) = E(f_i), i = 1, 2, \ldots, n\}
\]

Then, the conditional entropy of conditional probability distribution can be expressed by the following formula:

\[
H(P) = -\sum P(X)P(Y|X)\log P(Y|X).
\]

Among all models in set \(C\), the conditional probability model with the largest conditional entropy is called the maximum entropy model, and the logarithm in the above formula refers to the natural logarithm.

The advantage of the maximum entropy model is that when modeling training data, we only need to pay attention to the selection of features, instead of spending a lot of energy on how to make good use of these features.

### 3.3. Grammar Checking Based on ML Algorithm

As a general rule, grammar checking technology uses basic natural language processing technology, such as word segmentation, part-of-speech tagging, and corpus. Tag words in sentences based on grammar and context with part-of-speech tagging. Part-of-speech tagging technology has been widely used in all aspects of natural language processing because it has always been regarded as an important basic research field in natural language processing. The majority of today’s part-of-speech tagging systems are based on statistical or regular models. Mark each word’s corresponding part of speech as part of speech, such as verb, noun, adjective, adverb, or other part of speech, before checking the sentence grammar. To mark words, various corpora employ a variety of tag set conventions. The rule-based module and the statistics-based module are the two parts of the grammar checking module. To check the grammar of the input sentence, this document employs a method that combines the two. The architecture diagram of this module is shown in Figure 3.

Compared with the traditional GBDT (Gradient-Based Decision Tree) method, XGBoost (Extreme Gradient Boosting) is better in error approximation and numerical optimization. In recent years, XGBoost has become one of the most popular methods in various ML-based applications and competitions. The XGBoost model not only predicts better than the GBDT model but also trains faster than the GBDT model, which is why the XGBoost model is widely used by competitors in various data mining competitions.

Suppose there is a model composed of \(K\) trees:

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F.
\]

Solve the objective function of each parameter in the tree:

\[
\text{Loss} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)
\]

\[
= \sum_i (y_i - \hat{y}_i)^2 + \sum_k \left(yT + \frac{1}{2}\lambda\|w\|^2\right).
\]

Among them, \(\Omega(f_k)\) includes two parts: parameter \(y\) reflects the influence of the number of leaf nodes \(T\) on the error; \(\lambda\) parameter reflects the influence of leaf node weight \(w\) on error.

RF (random forest) is an extended variant of bagging. Through self-help resampling technology, \(n\) samples are randomly selected from the training set \(N\) to form a new training set to train the decision tree, and then \(n\) decision trees are generated to form RF. Its essence is the improvement of the decision tree algorithm. Each tree is built with independently selected samples, and these trees are merged together.

Averaging the trained \(K\) decision tree results,

\[
\text{Test}.y_i = \frac{1}{K} \sum_{k=1}^{K} T_h(\text{Test}_i).
\]

It is abstracted that if a query contains the keyword \(w_1, w_2, \ldots, w_n\), their word frequencies in a specific text are \(TF_1, TF_2, \ldots, TF_n\). The relevance of the query sentence to the text is \(\sum_{i=1}^{n} TF_i\).

Assuming that a keyword \(w\) has appeared in \(D_w\) sentences, the larger \(D_w\) is, the smaller the weight of \(w\) is, and vice versa. Let \(D\) be the total number of sentences. At present, the most used weight is “inverse text frequency index,” and its formula is

\[
\text{Index} = \log \frac{D}{D_w}
\]

The setting of this weight must meet the following two conditions: the stronger the ability of a word to express its main idea, the greater the weight, conversely, the smaller the weight. The weight of deleted words should be zero.
When evaluating the similarity between words, we use Spearman correlation coefficient as the evaluation index, which is a nonparametric index to measure the dependence of two variables. For the sample set with \( n \) number, \( X_i, Y_i \), respectively, represent two variable values of sample \( i \), and \( X, Y \) are sorted in increasing or decreasing order at the same time; \( x_i, y_i \), respectively, represent the rank of \( X_i, Y_i \) after sorting, and the formal definition of Spearman correlation coefficient \( \rho \) is

\[
\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
\]  

where \( d_i \) is the difference of variable ranking level, \( d_i = x_i - y_i \).

4. Experiment and Results

This paper selects the data set published by the international data mining platform composition scoring contest. The composition data set includes eight data subsets, each of which has a corresponding composition theme. Students write according to the description requirements of the composition theme. Each training data subset contains more than 1,000 students’ learned compositions and corresponding manual grading scores, and the number of words in each composition is mostly between 150 and 600. This document selects English Wikipedia corpus as the training corpus and uses Word2Vec to train the language model. The window size is set to 6, the word vector dimension is set to 500, and the truncated word frequency is set to 6. The number of cluster centers of vector clusters is set to 25. Based on nontext features and reasoning features, the content text features extracted based on word vector clustering are added to the model, and the training results are shown in Table 1.

As can be seen from Table 1, after adding the content text features based on word vector clustering, the effect of each model is greatly improved, especially the maximum entropy model, which is improved by 0.048. The XGBoost model has also been greatly improved, from 0.771 to 0.803, and its prediction effect has surpassed RF, which is the best among the three models. In addition, in general, CBOW is suitable for small training corpus, while Skip-Gram is more suitable for large corpus. The Wikipedia English corpus used in this paper exceeds 17G, so it is still a relatively large training corpus, so the performance of Skip-Gram is also better. After training several words to embed into word vectors, we group words according to their semantic information through clustering and then extract the features of items in each semantic category for model training. In the previous experiment, the number of cluster centers in the cluster was set to 30, and the influence of cluster centers on the experimental results is shown in Figure 4.

As can be seen from Figure 4, with the increase of the number of clustering centers, the experimental results of various word vector methods first increase slightly and then gradually decreased. When the cluster center is between 15 and 25, the scoring effect is the best, because when the number of initial clustering centers is small, the grouping restriction based on semantic information is weak, and the semantic information of each category is chaotic. As the number of clusters increases, the clustering constraints become stronger, and the semantic information of words is more consistent. When word embedding is used to represent a word as a vector of a certain dimension, each dimension of the word vector represents the implied semantic features of the word. Figure 5 compares the effects of word vectors of different dimensions on the experimental results.

It can be seen that with the increase of word vector dimension, the Skip-Gram effect and nontext features first increase and then decrease, but the change range is very small. However, CBOW, nontext features, and deduction features increase with the increase of word vector dimension, and the scoring effect shows irregular changes, but the change range is basically small. It can be seen that when the word embedding method is used to obtain word vectors, the dimension of word vectors has little influence on the semantic information of words. 100 English compositions are randomly selected from the test corpus as the test set, which contains 1228 sentences, of which the total number of marking errors is 1063. In this test set, the syntax checking module and results of this paper are shown in Table 2.

Similarly, 100 articles were randomly selected from the corpus as the test corpus, and the test set was checked for errors, and the accuracy rate was 68%. On the contrary, the accuracy of this model has been improved to some extent. For the evaluation indexes of malicious composition detection algorithm and composition detection algorithm unrelated to the subject to be tested, we use FPR (false positive rate), FNR (false negative rate), and correct rate which are commonly used abroad. Because the detection algorithm designed in this work has nothing to do with the object to be tested, it is necessary to classify the similarity between the composition to be tested and the object itself, and the similarity between the composition to be tested and the object to be tested. In addition, we analyze the similarity between the components to be tested and the subjects in the
The classification determines whether the composition is off topic, so the classification threshold must be determined by experiment. Some experimental results are shown in Figure 6.

It can be seen that when the classification threshold increases, the value of FPR will decrease and the value of FNR will increase correspondingly. When the off-topic threshold is set to 13, the correct algorithm of topic-independent composition detection rate has reached the maximum value of 90.12%. At this time, the probability of being off-topic is 1.03%, and the probability of being off-topic is 12.36%. In practical application, the classification threshold can be appropriately increased to further reduce the probability that digression is misjudged as digression composition. Although completely off-topic articles will be judged as off-topic, this type of composition will not get higher scores in extracting off-topic sentences and scoring algorithms. In order to verify the effectiveness of the composition detection algorithm that has nothing to do with the subject to be tested, this paper will compare it with the existing unsupervised detection algorithms outside the subject. The experimental results of the three algorithms are shown in Figure 7.

It can be seen from the experiment that the algorithm proposed in this paper is lower than the other two algorithms in FPR and FNR and has achieved good results [17].

| Table 1: Experimental results of content text features based on word vector clustering. |
|-------------------------------|----------------|----------------|
| Features                      | XGBoost        | RF             | Maximum entropy |
| Nontextual features           | 0.771          | 0.701          | 0.743           |
| Nontext features and deduction features | 0.793          | 0.796          | 0.781           |
| CBOW                          | 0.798          | 0.793          | 0.788           |
| Skip-Gram                     | 0.803          | 0.801          | 0.791           |

| Table 2: Check the test results of the grammar module on the test set. |
|--------------------------|----------------|----------------|
|                           | Precision      | Recall         | F value |
| Verb error detection     | 63.2           | 36.3           | 38.7   |
| Subject-predicate consistency error detection | 54.7           | 27.4           | 30.3   |
| Error detection of singular and plural nouns | 71.3           | 55.6           | 66.1   |
| Article error detection  | 77.2           | 41.8           | 54.8   |
| Error detection of prepositions | 59.6           | 39.7           | 50.6   |
| Overall error detection  | 88.7           | 60.1           | 44.8   |
| Ref [18]                 | 68.3           | 41.8           | 36.2   |

Reference [18]: English grammar error detection system for Chinese learners.
algorithm in this model is used to score the relevance of 700 articles. Finally, the scores of the algorithm are compared with those of two English teachers in the research group. The corresponding results are shown in Figure 8.

This model can not only detect the entire off-topic composition with high accuracy, but it can also score the composition’s relevance, and the result is very similar to the teacher’s score. Because college students’ English writing requirements are typically 120–150 words, the word count reflects the length of the composition. This function can determine if the composition is empty, too long, or too short. The author’s command of verbs is reflected in the number of verbs used. This feature has been studied by relevant scholars at an early stage, and results show that it has a high predictive ability for college English writing composition scores. A new model is built by adding statistical information, word grouping, word pronunciation, and other features to the basic feature set, and the new model is supervised and trained by manually labeling data, significantly improving the new model’s accuracy.

5. Conclusions

The intelligent evaluation model for English composition has the following benefits: high suitability, high evaluation, quick feedback, and low educational cost. It should be widely used because it plays an important role in the development of language skills and the enhancement of language ability in both mother tongue and second language learning. The method of extracting word diversity features from word vector clustering using machine learning is investigated in this paper. The experimental results show that by combining topic information, the word vector method can represent words as different word vectors under different topics, more accurately represent the semantic information of words in the text, and achieve the best effect on several text features. The effect of each model improves significantly after adding the content text function based on word vector clustering, particularly the maximum entropy model, which improves by 0.048. The XGBoost model has also been significantly improved, with a score of 0.803 compared to 0.771 previously. The test corpus is made up of 100 articles chosen at random from the corpus and checked for errors. It has a 68 percent accuracy rate. This model’s accuracy has gotten a little better. It demonstrates that the method used in this paper is precise and useful [21].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author does not have any possible conflicts of interest.
References


