Machine Learning Models: A Study of English Essay Text Content Feature Extraction and Automatic Scoring

Wei Shang*, Huihua Men and Xiujie Du

School of Humanities, Shandong Agriculture and Engineering University, Jinan, Shandong 250000, China
E-mail: weilun97839xls@yeah.net
**Corresponding Author*

Received 13 June 2023; Accepted 06 August 2023;
Publication 18 November 2023

Abstract

Accurate automatic scoring of English essay is beneficial for both teachers and students in English teaching. This paper briefly introduced an XGBoost-based automated scoring algorithm for English essay. To improve the accuracy of the algorithm, a long short-term memory (LSTM) semantic model was introduced to extract semantic scoring features from essays. Finally, the improved XGBoost algorithm was compared with the traditional XGBoost and LSTM algorithms in a simulation experiment using five types of essay prompts. The results indicate that the improved XGBoost algorithm has the best performance for automatic scoring of English essay and also requires the shortest scoring time.

Keywords: Machine learning, English essay, feature extraction, automatic scoring.

Journal of ICT Standardization, Vol. 11_4, 379–390.
doi: 10.13052/jicts2245-800X.1143

1 Introduction

English essay is a crucial part of English learning and reflects students' overall ability to apply English language skills. Therefore, accurate scoring of English essay is essential [1]. The traditional evaluation method involves manual grading by teachers, which is time-consuming and may not provide personalized evaluation for each student, especially when faced with a large number of students [2]. The emergence of machine learning algorithms provides a method for automatic scoring of English essay. The basic principle of machine learning algorithms for automatic essay scoring is to use a large dataset of pre-scored essays to train the algorithm to learn the scoring pattern and apply it to score unknown essays. Applying machine learning to automatic English essay scoring can make the scoring more objective and efficient, saving teachers' time and effort [3]. However, the use of machine learning for automaticessay scoring still has limitations in capturing subjective information such as writing style and context. The algorithm needs to be further improved to account for these subjective elements. Mcnamara [4] studied the application of the hierarchical classification method in automatic essay scoring and proved the validity of the method in the field of essay scoring. Li [5] proposed a new model for automatic Chinese essay scoring using a neural network, which applies the BERT network to obtain the sentence vector of an article and then extracts the article vectors using a two-layer bidirectional long short-term memory (Bi-LSTM) network. The experimental results showed that this model had better performance than other baseline methods. Hao [6] presented a weighted finite-state automaton-based system and utilized incremental latent semantic analysis to process massive essays. The experiment results verified the effectiveness of the system. This article briefly introduced an XGBoost-based automatic scoring algorithm for English essays and introduced an LSTM semantic model to extract semantic scoring features from essays to improve the accuracy of the algorithm. Finally, the optimized XGBoost algorithm was compared with the traditional XGBoost and LSTM algorithms in a simulation experiment using five types of topic-given essays.

2 Machine Learning Based English Essay Scoring Algorithm

The XGBoost algorithm is a common machine learning algorithm. The traditional automatic English essay scoring algorithm will first extract the shallow

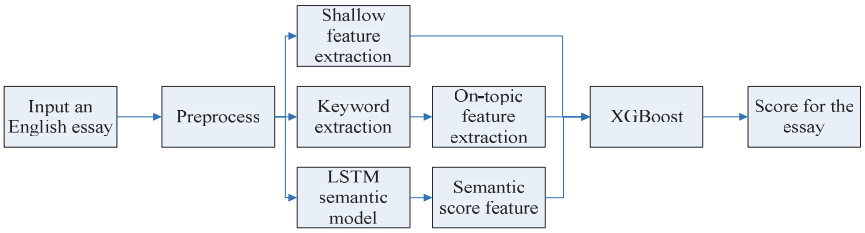


Figure 1 Basic process of machine learning-based english essay scoring algorithm after combining semantic information.

features and on-topic features of the essay when scoring [7]. The shallow features of the essay are the statistical features that only consider the structural organization without considering the meaning of the text. The word level includes the number of words, the number of word misspellings, the number of words after deduplication, the length of words, and the part of speech of words, etc. The sentence level includes the number of sentences, the length of sentences, the number of paragraphs, etc. The on-topic features are the keywords of the essay compared with the keywords of the essay prompt to measure the similarity between them [8]. The extracted features are then input to the XGBoost algorithm for score prediction. The formula used to calculate the essay scores in the XGBoost algorithm is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in F, \tag{1}$$

where $\hat{y}_i$ is the computed prediction score, f_k is the base learner, F is the set of base learners [9], totally K base learners, x_i is the extracted feature of the essay. The objective function [10] used in training the XGBoost algorithm is:

$$loss^{(t)} = \sum_{i=1}^{n} l(y_i, (\hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t), \tag{2}$$

where $loss^t$ is the current objective function (loss function), y_i is the actual score corresponding to x_i, $\hat{y}_i^{t-1}$ is the predicted score of the previous $t-1$-th integrated learner for x_i, $f_t(x_i)$ is the predicted score of the current learner for x_i, and $\Omega(f_t)$ is the regular term of the current learner.

In the above-mentioned traditional English essay automatic scoring algorithm, the shallow features reflect the good or bad structure of the essay, and the on-topic features reflect whether the topic of the essay fits the topic [11].

In this paper, the on-topic degree will be defined as semantic similarity between a topic given for an essay and the description of the topic in the essay. The level of similarity between these two factors is used to indicate the on-topic degree. However, in the actual manual scoring process, the reviewer will not only score from the above two types of features, but also pay attention to the semantic content of the essay as a whole, so the automatic scoring algorithm also needs to evaluate the semantics of the essay. The automatic scoring algorithm also needs to evaluate the semantics of the essay. This paper introduces a semantic model into the traditional automatic English essay scoring algorithm to obtain the semantic scoring features of the essay, so as to improve the scoring accuracy of the algorithm.

(1) An English essay is input and pre-processed by cleaning noisy data, converting letters uniformly to lowercase, replacing irregular symbols, etc. [12].
(2) The shallow features of the essay are extracted, and the shallow feature items to be extracted are shown in Table 1. Sentence readability is the weighted sum of the average number of characters per word and the average length of the sentence, which reflects the difficulty in reading.
(3) The keywords are extracted. First, the stop words are removed. Then, the number of occurrences of each word in the text is counted, and the term frequency-inverse document frequency (TF-IDF) value of each word is calculated [13], and the top 5 words with the largest TF-IDF values are considered as keywords. The keywords of the essay and the keywords of the essay prompts are converted into word vectors using the

Table 1 Shallow features

Feature Level	Feature Name	Feature Number
Word level	Number of misspelled words	W_1
	Preposition usage ratio	W_2
	Number of connecting words	W_3
	Ratio of CET4 vocabulary	W_4
	Ratio of CET6 vocabulary	W_5
	Total number of words	W_6
	Average word length	W_7
	Word length variance	W_8
Sentence level	Number of sentences with grammatical errors	S_1
	Total number of sentences	S_2
	Average sentence length	S_3
	Sentence readability	S_4

Word2vec model. The cosine distance mean value between the keywords of the essay and the keywords of the prompt is calculated using the word vectors as the on-top feature.

(4) The semantic score of the essay is computed using the LSTM semantic model. First, the essay is converted into word vectors using the Word2vec model [14], and then hidden state h_t of the essay is computed in the hidden layer of the LSTM:

$$h_t = LSTM(sent_1, sent_2, \ldots sent_t), \tag{3}$$

where $sent_t$ is the Word2vec word vector of the t-th sentence of the English essay. Finally, h_t gets a score between 0 and 1 in the fully connected layer by the sigmoid function, which is the semantic score of the essay [15].

(5) The shallow features, on-topic features, and semantic score features of the English essay are input into the XGBoost algorithm, and the predicted score of the essay is calculated according to Equation (1) of the XGBoost algorithm.

3 Simulation Experiments

3.1 Experimental Data

The English essay scoring data needed to conduct the simulation experiment were obtained from the essays of freshman and sophomore students in the midterm and final English exams at Shandong Agriculture And Engineering University. Five thousand essays were selected as the dataset. There were five prompts in the dataset, and the number of essays, the lowest (highest) score, the average score, and the average number of words in each prompt are shown in Table 2. The data in Table 2 showed that there were low-scoring

Table 2 Overview of the English essay data set

	Number of Essays	Lowest Score	Highest Score	Average Score	Average Word Count
Prompt 1	800	2	33	25	315
Prompt 2	1100	0	42	31	310
Prompt 3	900	0	47	30	300
Prompt 4	1200	3	50	32	312
Prompt 5	1000	2	44	29	300

essays with scores of 0, 2, and 3 points among the collected essays. These types of essays are also necessary for two reasons: firstly, automatic scoring algorithms require a sufficient range of scores for training to ensure accuracy; secondly, low-scoring essays can reflect more typical problems and have evaluative value as well. Then, 60% of the essays from every prompt were randomly selected as the training set and the remaining 40% as the test set.

3.2 Experimental Setup

In the automatic scoring algorithm, the third-party Python tools SpellCheck and nltk were used to check the spelling and segmentation of words, and the number of prepositions, conjunctions, and CET-4 and CET-6 words were obtained by a comparison with a vocabulary list. When extracting the on-topic feature, the vector dimension of the Word2vec model, which converted keywords into word vectors, was set to 250. When calculating the semantic score of essays using the LSTM semantic model, 64 neurons were set in the LSTM hidden layer, and the activation function was the sigmoid function. In the XGBoost algorithm, the base learner for iterative training was a linear model, and the learning task was linear regression between the input features and the essay score. The threshold for node splitting during training was used to determine whether the base leaner is split or not. If splitting can make the reduction of the loss function larger than the threshold, then it is split. After an orthogonal experiment, the threshold for node splitting was set to 0.1, and the learning rate was set to 0.05.

To further verify the performance of the automatic scoring algorithm, it was compared with two other algorithms. The first one was the traditional XGBoost scoring method, which did not introduce the semantic scoring model and used only shallow features and on-topic features as inputs to the XGBoost algorithm. The second scoring algorithm used only the LSTM algorithm to score the essays and also used 250-dimensional Word2vec word vectors as inputs. The LSTM algorithm, derived from RNN, takes into account the influence of context when processing data, making it particularly suitable for handling sequential data such as essays.

3.3 Evaluation Criteria

In this paper, the Kappa value was used to measure the performance of the scoring algorithm, The scoring of essays was set to N levels. In the dataset, essays were scored in the range from 0 to 60, all integers, so $N = 61$.

The formula for calculating Kappa is:

$$\begin{cases} \omega_{ij} = \left(\dfrac{i-j}{N-1}\right)^2 \\ k = 1 - \dfrac{\sum_{ij} \omega_{ij} O_{ij}}{\sum_{ij} \omega_{ij} E_{ij}} \\ z = \dfrac{1}{2} \ln \dfrac{1+k}{1-k} \\ kappa = \dfrac{e^{2z}-1}{e^{2z}+1} \end{cases}, \tag{4}$$

where O_{ij} denotes the number of essays manually scored as i and algorithmically scored as j, E_{ij} is the outer product of O_{ij}, ω_{ij} denotes the degree of difference between manual score i and algorithm score j as the weight of the corresponding position thereafter, k is the quadratically weighted Kappa value, and z is the value after Fisher transformation of k.

3.4 Experimental Results

The automatic English essay scoring algorithm used in this paper requires the extraction of relevant features of the essay and then the calculation of the essay score by the XGBoost algorithm. The essay features used in this process can all be expressed in the form of numerical values, and Table 3 shows the extracted features of some essays and their scores. From a simple comparison of the data in Table 3, it was initially found that the higher the essay score, the lower the number of incorrect words and sentences, and the higher the percentage of advanced vocabulary, on-topic degree, and semantic score in the essay.

Figure 2 shows the kappa values of three automated essay scoring algorithms for five types of essay prompts. The specific values are shown in the data labels in Figure 2. It was seen from the figure that regardless of the type of essay prompt, the improved XGBoost algorithm had the highest kappa value for scoring English essays, the LSTM algorithm had the second-highest kappa value, and the traditional XGBoost algorithm had the lowest value. The reason for this is that the traditional XGBoost algorithm uses only shallow features and on-topic features, which only reflects whether the essay structure is standardized and whether the essay is on-topic. In general, well-organized and on-topic essays will not have a poor score. However,

Table 3 Extracted features of some essays and their scores

Features	No. 1 Essay (Prompt 1)	No. 15 Essay (Prompt 2)	No. 98 Essay (Prompt 3)	No. 136 Essay (Prompt 4)
Number of misspelled words	5	10	3	1
Preposition usage ratio	0.43	0.21	0.56	0.67
Number of connecting words	34	12	54	62
Ratio of CET4 vocabulary	0.22	0.12	0.25	0.29
Ratio of CET6 vocabulary	0.21	0.11	0.26	0.31
Total number of words	421	432	411	429
Average word length	9	8	11	13
Word length variance	1.1	3.2	0.9	0.8
Number of sentences with grammatical errors	4	6	3	2
Total number of sentences	30	30	31	32
Average sentence length	10	8	11	13
Sentence readability	0.7	0.5	0.8	0.9
On-topic degree	0.6	0.3	0.7	0.8
Semantic score	0.73	0.67	0.78	0.83
Score	25	15	36	44

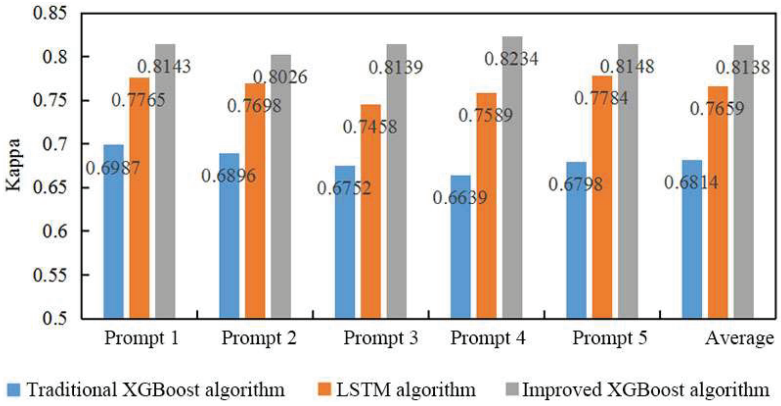


Figure 2 Kappa values of three automatic English essay scoring algorithms.

in the actual scoring process, the expression of the essay content is also an important factor and the main basis for scoring. The LSTM algorithm can connect the context and evaluate the expression of the essay content, but it also lacks the evaluation of the essay structure and the degree of off-topic. For an essay, no matter how well the content is expressed, if it is off-topic,

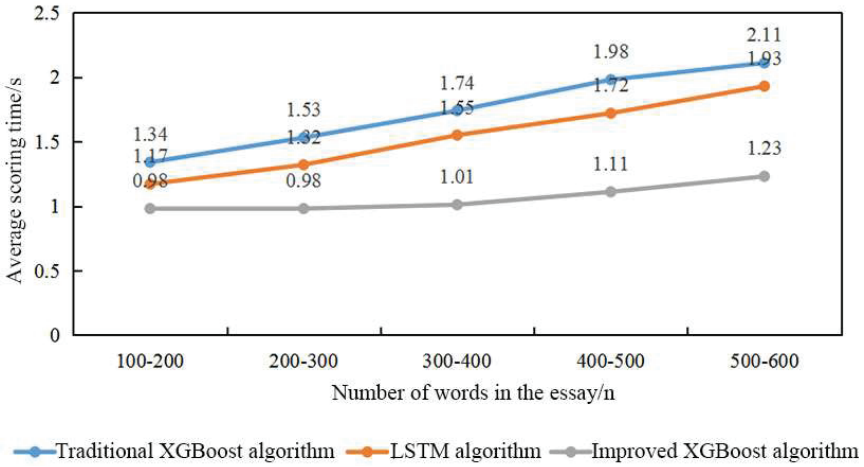


Figure 3 Average scoring time of three automatic English essay scoring algorithms for essays of different lengths.

it cannot get a high score. Therefore, the improved XGBoost algorithm combines the above two algorithms, referring to the shallow features representing the writing structure, the on-topic features representing the on-topic degree, and the semantic score features representing the content quality, resulting in the best scoring performance.

Figure 3 shows the average scoring time of three automatic essay scoring algorithms for essays of different lengths, with specific values shown in the data labels in the figure. It was seen that as the number of words in the essay increased, the scoring time of all three automatic scoring algorithms also increased. This is because an increase in the number of words leads to an increase in the amount of data that the algorithm has to process. Among them, the time increase for the improved XGBoost algorithm was relatively small. Under the same range of essay lengths, the traditional XGBoost algorithm had the longest average evaluation time, followed by the LSTM algorithm, and the improved XGBoost algorithm had the shortest time.

4 Conclusion

This article briefly introduced an XGBoost-based automatic scoring algorithm for English essays. To enhance the accuracy of the algorithm, an LSTM semantic model was introduced to extract semantic scoring features from essays. Finally, the improved XGBoost algorithm was compared with

traditional XGBoost and LSTM algorithms in a simulation experiment using five types of essay prompts. The final results are as follows. (1) The higher the essay score, the fewer errors in words and sentences, the higher the proportion of advanced vocabulary, on-topic degree, and semantic score in the essay. (2) Regardless of the type of essay prompt, the improved XGBoost algorithm had the highest kappa value for scoring English essays, the LSTM algorithm had the second-highest kappa value, and the traditional XGBoost algorithm had the lowest value. (3) As the number of words in the essay increased, the scoring time for all three automatic scoring algorithms increased; under the same range of essay lengths, the traditional XGBoost algorithm had the longest average scoring time, followed by the LSTM algorithm, and the improved XGBoost algorithm had the shortest time.

References

[1] B. Uysal, N. Doan, 'Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test', Int. J. Assess. Tools Educ., 8(2):222–238, 2021.

[2] M. Beseiso, S. Alzahrani, 'An Empirical Analysis of BERT Embedding for Automated Essay Scoring', Int. J. Adv. Comput. Sc., 11(10): 204–210, 2020.

[3] L. Xia, A. Mc, C. Jyn, 'SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring', Knowl.-Based Syst., 210:106491.1–106491.14, 2020.

[4] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, J. Dai, 'A hierarchical classification approach to automated essay scoring', Assess. Writ., 23:35–59, 2015.

[5] H. Li, T. Dai, 'Explore Deep Learning for Chinese Essay Automated Scoring', J. Phys. Conf. Ser., 1631:012036, 2020.

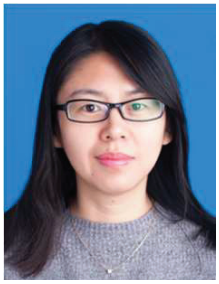
[6] S. Hao, Y. Xu, D. Ke, K. Su, H. Peng, 'SCESS: a WFSA-based automated simplified chinese essay scoring system with incremental latent semantic analysis', Nat. Lang. Eng., 22(2):291–319, 2016.

[7] T. C. Stephen, M. C. Gierl, S. King, 'Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation', Nurse Educ. Pract., 54(1):103085, 2021.

[8] L. Qian, Y. Zhao, Y. Cheng, 'Evaluating China's Automated Essay Scoring System iWrite', J. Educ. Comput. Res., 2019(2):073563311988147, 2019.

[9] R. Ahmad, 'E-learning Automated Essay Scoring System Menggunakan Metode Searching Text Similarity Matching Text', Jurnal Penelitian Enjiniring, 22(1):38–43, 2019.
[10] J. Contreras, S. Hilles, Z. B. Abubakar, 'Automated essay scoring using ontology generator and natural language processing with question generator based on blooms taxonomy's cognitive level', Int. J. Adv. Technol. Eng. Explor., 2019:2249–8958, 2019.
[11] K. Rysová, M. Rysová, M. Novák, J. Mírovský, E. Hajičová, 'EVALD – a Pioneer Application for Automated Essay Scoring in Czech', Prague Bull. Math. Ling., 113(1):9–30, 2019.
[12] H. M. Wang, 'Automated Chinese Essay Scoring Based on Deep Learning', Comput. Mater. Con., 2020(10):817–833, 2020.
[13] M. Yamamoto, N. Umemura, H. Kawano, 'Proposal of Japanese Vocabulary Difficulty Level Dictionaries for Automated Essay Scoring Support System Using Rubric', J. Oper. Res. Soc. China, 8(4):601–617, 2020.
[14] B. Uysal, N. Doan, 'Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test', Int. J. Assess. Tools E., 8(2):222–238, 2021.
[15] M. Beseiso, S. Alzahrani, 'An Empirical Analysis of BERT Embedding for Automated Essay Scoring', Int. J. Adv. Comput. Sc., 11(10): 204–210, 2020.

Biographies



Wei Shang, female, graduated from Shandong Normal University, majoring in Curriculum and Teaching of English, and received Master's Degree. Now she is working as a lecturer in School of Humanities in Shandong Agriculture and Engineering University, specializing in research of English teaching and

applied linguistics. She has participated in three provincial scientific research projects and published more than ten papers.

Huihua Men is a associate professor in the International Exchange and Cooperation Department in Shandong Agriculture and Engineering University, China. Her research interest include core competency, learning evaluation and English teaching. She has published more than 15 papers.

Xiujie Du graduated from Shandong Normal University, majoring in curriculum and teaching of English, and received Master's Degree. Now she is working as a lecturer in School of Humanities in Shandong Agriculture and Engineering University, specializing in research of English teaching and translation.