Journal of Network Intelligence
Taiwan Ubiquitous Information

©2023 ISSN 2414-8105 (Online)
Volume **8**, Number **1**, February **2023**

Multi-level Feature Fusion for Automated Essay Scoring

Jinshui Wang[1,2], Junyan Chen[1,2], Xuewen Ou[1,2], Qingfeng Han[3], Zhengyi Tang[1,2,*]

[1]School of Computer Science and Mathematics,
Fujian University of Technology, Fuzhou, 350118, China
[2]Fujian Provincial Key Laboratory of Big Data Mining and Applications,
Fujian University of Technology, Fuzhou, 350118, China
[3]China National Petroleum Corporation
wangjinshui@fjut.edu.cn,15866676515@163.com,954624549@qq.com,hqf@cnpc.com.cn,tangzy@fjut.edu.cn

*Corresponding author: Zhengyi Tang (tangzy@fjut.edu.cn)

Received July 6, 2022, revised August 18, 2022, accepted September 29, 2022.

Abstract. *Automatic Essay Scoring (AES) is one of the significant and challenging research topics in the Natural Language Processing(NLP) area. However, existing AES models majorly consider features derived from vocabulary while failing to integrate sentence and chapter information from multi-level perspectives. In this study, we proposed a multi-level feature fusion model, which was used to capture the multi-level features from different perspectives to improve the accuracy of scoring. Our model consisted of three components to respectively capture the vocabulary-level, sentence-level, and chapter-level features, which were then fed into a CNN and BiLSTM network for the final essay scoring. The results show that the proposed model outperforms a set of state-of-the-art AES models on the dataset of the Kaggle Automated Student Assessment Prize (ASAP) competition. The average Quadratic Weighted Kappa (QWK) value reaches 0.816, which verifies the efficacy of the model in the task of automated essay scoring.*

Keywords: Automated Essay Scoring, Feature Fusion, Natural Language Processing, BiLSTM, CNN

1. **Introduction.** AES is a natural language processing technology that uses linguistics, statistics and artificial intelligence to automatically score essays. It has become one of the hot research topics in the field of education. Since it was proposed in the middle of the last century [1], AES has been gradually applied to some large-scale language tests, such as the College English Test (CET) and the Graduate Record Examinations (GRE). AES can make up for the shortcomings of manual grading from the perspectives of teachers and students. On the one hand, AES can avoid the influence of human factors on the scoring results and reduce the marking workload; On the other hand, AES can enable students to get effective feedback timely and recognize their gaps.

The scoring results are influenced by various features. By sorting through previous studies, essay features can be majorly divided into the following categories: length-based features, readability features, embeddings features, part of speech features. The core idea of length-based features is that writing is usually required to be completed within a specified time. Therefore, counting the number of words, characters, etc., in the essay text effectively assesses the author's writing skills and thus reflects the quality of the essay. Readability features use a quantitative approach to evaluate the coherence between sentences and the ease of reading the utterance itself. Embeddings features are usually based on word vector embedding representations, and the extraction of features is performed using neural networks. For example, Beseiso et al. [2] used the RoBERTa language model combined with BiLSTM to score essays, respectively. Wangkriangkri et al. [3] explored Global Vectors for Word Representation (GloVe), Embeddings from Language Models (ELMo), and Bidirectional Encoder Representations from Transformers (BERT) and combined them with LSTM for AES, and compared the effectiveness of combining various word vector embeddings and neural networks. Part of speech features mainly reflect the standard of the essay by the number of different part of speech words in the essay [4]. In an essay, many nouns, verbs, adjectives, and adverbs are used, and the statements of the essay tend to be more vivid and full. The number of nouns, verbs, adjectives, and adverbs in the essay is counted, and the number of words in the part of speech category is used as a feature to measure the quality of the essay.

Only using the features of the above categories cannot comprehensively evaluate the pros and cons of the essay. In the actual scoring process, the basis for the essay score is mainly given around the three levels of vocabulary, sentence and chapter [5]. Therefore, it is crucial to use different levels of features to score essays. According to the different meanings of essay features, Essay features can be grouped into three levels: vocabulary-level features, sentence-level features, and chapter-level features. For example, length-based features and part of speech features can measure the vocabulary richness of the essay, so these two types of features are grouped into the vocabulary-level features. Readability features mainly quantify the articulation between sentences and the ease of reading the sentences themselves. Therefore, the readability features are grouped into the sentence-level features.

This study proposes a multi-level feature fusion AES model. The model uses CNN to extract local information at the vocabulary-level and sentence-level, and uses BiLSTM to extract global information at the sentence-level and chapter-level, and integrates the above information to predict essay scores. The contributions of this study are as follows.

- From a multi-level perspective, CNN and BiLSTM are used to extract the local and global information of the essay from different levels of features, respectively.
- A descriptive analysis of the different levels of features is conducted to analyze in detail their impact on essay scoring results.
- Experiment results demonstrate that our model achieves the best average QWK value and outperforms other state-of-the-art AES models almost on each subset.

2. **Related Work.** In recent years, with the increasing maturity of natural language processing technology, AES has become a popular direction of research in this field. According to the different approaches to achieving AES, this section reviews the available research results from the perspectives of machine learning, deep learning, and approaches that integrate feature engineering and deep learning.

In machine learning area, research on the AES has focused on feature engineering-based machine learning models [6, 7, 8, 9, 10]. Peter et al. [11] considered AES as a linear regression problem, and four categories of vocabulary-level manual features were extracted from the essays to complete the scoring of the essays by Bayesian Ridge Regression (BRR). The validity of the vocabulary-level features was experimentally demonstrated. Mahana et al. [12] extracted several features (total number of words, number of sentences, etc.) from the essays. They used not only a linear regression model to learn these features but also a Forward Feature Selection Algorithm (FFSA) to obtain the best-rated feature combinations. To investigate the effect of different features on scoring results, Zesch et al. [4] classified essay features into strong and weak dependency group features and used supervised machine learning models to learn the features, and found that features of the same type worked better when trained together. Chen et al. [13] used Listwise Learning to Rank algorithm while combining linguistic and statistical features for AES. Trung et al. [14] proposed to use Bayesian algorithm and MapReduce model to predict students' learning ability according to their academic achievements.

With the rapid development of deep learning, deep learning methods based on neural networks have achieved many research results in many research fields [15, 16]. In the field of AES, researchers usually use word embeddings combined with neural networks to score essays [17, 18]. Alikaniotis et al. [19] used a combination of word vectors and multiple neural networks to score essays, and then compared and analyzed the scoring results and finally found that the best results were achieved based on the BiLSTM model. Dong et al. [20] constructed a hierarchical CNN model with an upper layer representing the essay structure based on sentences representation and a lower layer representing the sentence structure, by which some high-level and abstract information of the essay was extracted. Tay et al. [21] proposed a new SkipFlow mechanism that could effectively alleviate the lack of LSTM memory capacity for learning discourse coherence of long texts by modeling.

Although many neural network-based scoring models have achieved advanced results, these models can be further improved by merging manual features obtained through feature engineering [22]. Liu et al. [23] developed a Two-Stage Learning Framework (TSLF), which combines manual features and deep coding features for AES. To effectively combine manual features with word vector-based text representations, Dasgupta et al. [24] proposed that using neural networks can augment manual features and achieve AES by using Convolution Recurrent Neural Network to train manual features and text representations separately.

As more and more features are used for the training of scoring models, it is inevitable that features of different properties (i.e. local and global) appear simultaneously in the same model. Due to the different data characteristics applicable to different neural networks, the scoring model using only a single neural network cannot simultaneously train the influence of local and global features of the essay on the scoring results. It is found that when marking essays, teachers' attention to vocabulary, sentence, and chapter in essays is cascading. Therefore, this study focuses on the different data features applicable to different neural networks and proposes a multi-level automatic scoring model for essays based on vocabulary, sentence, and chapter from the perspective of the actual assessment of essays by teachers.

3. Multi-level Feature Fusion for Automated Essay Scoring.

As shown in Figure 1, this study proposes a multi-level feature fusion AES model. At the vocabulary-level, we extract features representing the richness of the vocabulary used in the essay, and use CNN to obtain the local information of the vocabulary. At the sentence-level, for the readability features between sentences, local information of sentences is obtained through CNN. The Word2Vec is used to obtain the sentence representation, and the global information of the sentence is obtained by BiLSTM. At the chapter-level, the topic vectors are extracted using Latent Dirichlet Allocation (LDA), the semantic vectors are extracted using Latent Semantic Analysis (LSA), and the composition vectors are extracted using Doc2Vec. The global information at the chapter-level is obtained using BiLSTM after combining the above three. Finally, the above three levels of information are blended to predict the essay's score.

3.1. Extraction of vocabulary-level features.

The quality of a student's essay is closely related to the vocabulary in the essay. The vocabulary-level features can reflect the language skills of the students to a large extent and can also reflect the richness of

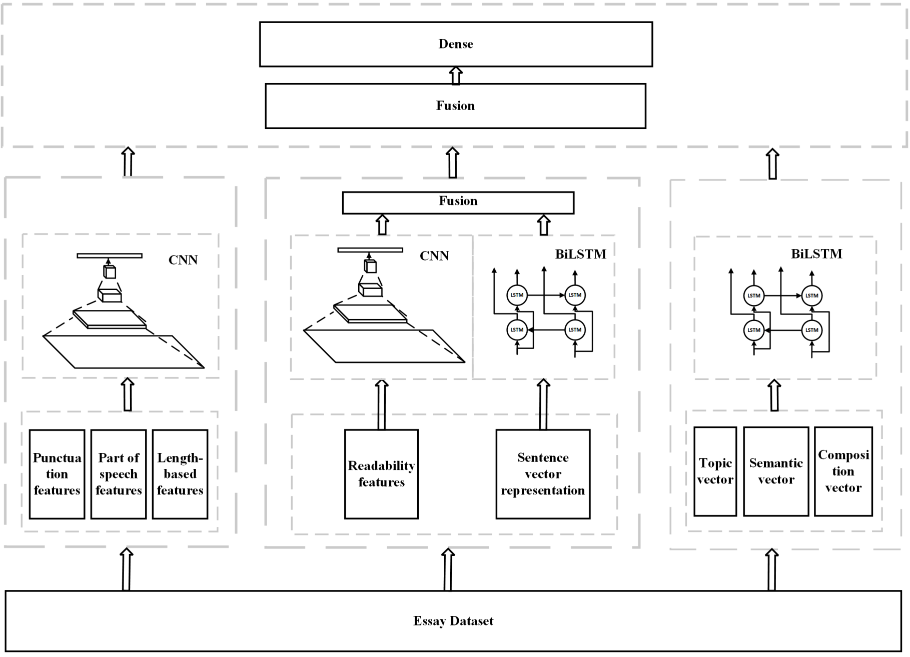

FIGURE 1. The architectures of the multi-level feature fusion framework

the essay content to a certain extent [11]. The selection of appropriate vocabulary-level features is an important factor that affect the performance of the essay scoring model [13]. The vocabulary-level features are shown in Table 1.

Length-based features are one of the most important feature types for AES, as length is positively correlated with an essay's score [22, 25]. These features code the length of an essay based on the number of words or characters in the essay.

Part of speech features contain a number of various vocabulary forms such as nouns, verbs, adjectives, adverbs, etc. These features are good indicators for testing vocabulary. Likewise, they can be used as basic indicators for phrasing [12].

Punctuation features are essential indicators of a well-structured and well-organized essay [12, 26]. The use of various punctuation marks can help express the author's thoughts and feelings.

TABLE 1. Vocabulary-level features

Feature Type	Description	Examples
Length-based features	the basic structure of essay length	'word_count','token_count', 'unique_token_count','nostop_count'
Part of speech features	information on the richness or diversity of students' vocabulary	'noun','adj','pron','verb','cconj','adv', 'det','propn','num','intj','ner_count'
Punctuation features	extraction of punctuation marks for essays	'comma','question','exclamation', 'quotation'

3.2. **Extraction of sentence-level features.** The score of the essay is the result of a multi-faceted comprehensive measure. Not only the vocabulary usage of the essay needs to be considered, but the sentence-level features of the essay are also an important aspect. Sentence-level features are mainly reflected in the word vector representation and readability features.

For the word vector representation of sentences, firstly, the essay is divided into sentences. Secondly, the Word2Vec model is used to sum and average the word vectors corresponding to each word in the sentence to represent the essay content. Suppose the word vector of word w is $f(w)$, then:

$$G(d) = \frac{1}{n}\sum_{i=1}^{n} f(w_i) \tag{1}$$

where $G(d)$ is used to represent the vector of essay d, n is the number of words contained in d.

Using the Word2Vec model to construct sentence text, this treatment only considers the average semantics of the sentences as a whole. It ignores the effect of the articulation between sentences on the semantics of the text. Therefore, the above problems can be effectively remedied by adding readability features. So this study adds readability features to Word2Vec to construct sentence text representations.

Readability features indicate the difficulty of the essay reading [27]. In a good essay, the writer should demonstrate various sentence structures and excellent cohesion between sentences [22]. Readability features are usually measured using four types of readability criteria: Readability_grades, Sentence_info, Word_usage, and Sentence_beginnings. The readability features are shown in Table 2.

TABLE 2. Readability features

Feature Type	Description	Examples
Readability_grades	readability index	'Kincaid','ARI','Coleman_Liau','LIX', 'FleschReadingEase','FleschReadingEase', 'GunningFog','SMOG','RIX','DaleChall'
Sentence_info	information about the sentence	'characters_word','syll_word','wordtypes', 'words_sentence','words','sentences', 'sentences_paragraph','complex_words', 'type_token','characters','syllables', 'paragraphs','long_words','complex_dc'
Word_usage	lexical features on sentence coherence indicators	'tobeverb','auxverb','conjunction', 'pronoun','preposition','nominalization'
Sentence_beginnings	sentence-to-sentence articulation index	'noun','interrogative','paper','preposition', 'subordination','conjunction'

3.3. **Extraction of chapter-level features.** The LDA topic model can give the topic of each essay in the form of a probability distribution, which cannot represent the features of the entire chapter-level [28]. To solve this problem, this study integrates the three models of LDA, LSA, and Doc2Vec to extract the chapter-level features of essays more comprehensively.

The topic probability distribution of the text can be obtained based on the LDA topic model, and it can be extracted as the topic features of the text. Given the set of essays $D = d_1, d_2, ..., d_n$, in mining the topics $t_1, t_2, ..., t_m$, where the probability distribution of any word $w_{i,j}$ under topic t_i is $\theta_{i,j}$, the top α words with the highest probability under each topic are selected, and the probability distribution $\lambda_{i,j}$ of the words is recalculated according to Equation (2), based on which the topic vector $v(t_1), ..., v(t_q)$ can be obtained by Equation (3).

$$\lambda_{i,j} = \frac{\theta_{i,j}}{\sum_1^{\alpha} \theta_{i,j}} \tag{2}$$

$$v(t_i) = \sum_{j=1}^{\alpha} \lambda_{i,j} v(w_{i,j}) \tag{3}$$

LSA maps words and chapters to the latent semantic space, thus removing some "noise" from the original vector space and improving the accuracy of information retrieval. A large collection of essays is statistically analyzed to extract the meaning of the words in the context of use. Technically, the influence of synonyms and polysemous words is eliminated by processing such as SVD decomposition, and the accuracy of subsequent processing is improved. LSA focuses on mining the contextual semantics of chapters by decomposing the matrix between words and texts. The word-text matrix is represented using the product of the word-topic matrix T and the topic-text matrix Y. The word-text matrix can be represented by Equation (4).

$$v(x_i) = T_i \times Y_i \tag{4}$$

LDA and LSA models can mine the relationship between topic and context in a chapter from the global but ignore the semantic information of the chapter as a whole. In contrast, the Doc2Vec model compensate for the above disadvantages. The essay set $D = d_1, d_2, ..., d_n$ is trained under the Doc2Vec model to obtain the word vectors $v(w_1), ..., v(w_n)$ and the paragraph vectors $v(p_{1,1}), ..., v(p_{n,rn})$, which $w_1, w_2, ..., w_N$ represents the words in the vocabulary and $p_{1,1}, ..., p_{n,rn}$ represents the chapter paragraphs. Then the composition vectors are calculated by Equation (5), where s is the number of paragraphs in the essay text.

$$v(d_i) = \frac{\sum_{j=1}^{s} p_{i,j}}{s} \tag{5}$$

Finally, the topic vectors, the semantic vectors, and the composition vectors are used as the chapter-level features.

3.4. **CNN and BiLSTM model.** CNN can be used both to extract relationships between local features and to mine new feature information through fusion between features to enhance the original features. Based on the above advantages of CNN, this study uses CNN to train vocabulary-level features with sentence-level readability features. CNN has two main operations: convolution and pooling. The convolution layer uses a number of sliding windows of size $k \times h$ to convolve the input essay feature matrix from different levels, i.e.

$$c_i = f(w \cdot X_{i:i+h-1} + b) \tag{6}$$

where c_i denotes the i eigenvalue, $f(\cdot)$ is the convolution kernel function, $w \in R^{hk}$ is the filter, h is the sliding window size, and b is the bias value. $X_{i:i+h-1}$ denotes the local eigenmatrix consisting of the i row to the $i + h - 1$ row.

In the pooling layer, this study uses the maximum pooling method to take the maximum value of each vector of the output of the convolutional layer to extract the most important essay feature information and then connect them into a vector to get the output of the pooling layer. The maximum pooling approach automatically extracts the most useful essay features using neural networks.

$$C = [c_1, c_2, c_3 ..., c_{n-h+1}] \tag{7}$$

$$\hat{c} = max\,\{C\} \tag{8}$$

Among the current deep models, the LSTM model is widely used in various fields because of its ability to effectively utilize long-distance dependency information in sequence data [29]. The advantage of BiLSTM is that bi-directional LSTM can use both past moment and future moment essay information and predict the final essay score more accurately than one-way LSTM. In the vector representation of sentence-level and chapter-level features, the bi-directional LSTM model can capture the complete semantic information at the sentence-level and chapter-level by combining forward-layer LSTM and backward-layer LSTM. Therefore, for the features of the semantic level of sentences and chapters, this study chooses BiLSTM training. In the LSTM model, each neuron contains three gates, namely the forgetting gate, input gate and output gate, to protect and control the information state [30, 31, 32]. The calculation formula is as follows.

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{9}$$

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \tag{10}$$

$$\widetilde{c}_t = tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c) \tag{11}$$

$$c_t = i_t \circ \widetilde{c}_t + f_t \circ c_{t-1} \tag{12}$$

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{13}$$

$$h_t = o_t \circ tanh(c_t) \tag{14}$$

where x_t is the input vector at moment t. W_i, W_f, W_c, W_o and U_i, U_f, U_c, U_o are the weight matrices, and b_i, b_f, b_c, b_o are the bias terms. The symbol $\circ$ denotes the matrix algorithm, and σ denotes the *sigmoid* function.

In AES, to fully use the contextual information of sentence and chapter texts, BiLSTM would be used, a combination of two LSTM models with opposite temporal order.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t) \tag{15}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t) \tag{16}$$

$$H_t = \langle \overrightarrow{h_t}, \overleftarrow{h_t} \rangle \tag{17}$$

where $\vec{h}_t$ and $\overset{\leftarrow}{h}_t$ are hidden layers, $\vec{h}_t$ aims to obtain the forward essay semantic information by forward LSTM, and $\overset{\leftarrow}{h}_t$ aims to obtain the reverse essay semantic information by backward LSTM. H_t is the semantic feature of the essay.

4. **Experimental setup.** In this study, we used the publicly available dataset of the Essay Scoring Contest on Kaggle, an international data mining platform, and used 5-fold cross validation on the ASAP training data for evaluation. The specific information of the essay set is shown in Table 3.

The quadratic weighted kappa (QWK) chosen in this study to evaluate the performance of the model was a consistency test to assess whether the scores derived from the model were consistent with the actual scores. Assuming that the score of the essay can be divided into N grades, the calculation formula of QWK is as follows:

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{18}$$

$$w_{i,j} = \frac{(i-1)^2}{(N-1)^2} \tag{19}$$

Table 3. Statistical results based on the essay dataset

Essay	Grade	Number	Word count	Score range	Genre
1	8	1783	350	0-12	Explanatory Essay
2	10	1800	350	0-6	Explanatory Essay
3	10	1726	150	0-3	Material Essay
4	10	1772	150	0-3	Material Essay
5	8	1805	150	0-4	Material Essay
6	10	1800	600	0-4	Material Essay
7	7	1569	250	0-30	Explanatory Essay
8	10	723	650	0-60	Explanatory Essay

where O is the n order histogram matrix, $O_{i,j}$ denotes the number of essays with actual score i and model score j, $W_{i,j}$ denotes the quadratic weighting matrix based on the difference between actual and model scores, and $E_{i,j}$ denotes the product of the probabilities of having actual score i and model score j.

In the model training process, in order to improve the model prediction performance and generalization ability, after many experiments and debugging, the parameters shown in Table 4 are selected.

Table 4. Model parameters

Level	Parameter name	Parameter value
vocabulary-level	CNN hidden unit dimension	64
sentence-level	CNN hidden unit dimension	64
	BiLSTM hidden unit dimension	64
chapter-level	BiLSTM hidden unit dimension	128
	Dropout rate	0.2
	Epochs	200
	Batch size	128

5. **Baseline methodology and implementation details.** H1-H2 represented the results scored by 2 manual raters.

BLRR [11] modelled the automatic essay scoring task as a regression problem and used Bayesian Ridge Regression as the learning algorithm.

CNN+LSTM [33] proposed a model that did not require any feature engineering but was based on recurrent neural networks to learn the relationship between an essay and was specified score.

LSTM-CNN-att [34] built a hierarchical sentence-document model to represent essays, using attention mechanisms and neural networks to automatically determine the relative weights of words and sentences and to learn essay representations.

RL1[35] used a reinforcement learning model that used a classification approach for scoring.

SKIPFlOW [21] proposed a unified deep learning architecture to generate neural coherence features in an end-to-end manner.

HISK+BOSWE [36] proposed an automatic scoring model for essays based on a combination of string kernels and word embeddings, where the string kernels captured similarities between strings based on the counting of common characters n-gram.

TSLF-ALL [23] proposed a two-stage learning framework (TSLF), which utilized deep encoding features and manual features.

R^2BERT [37] utilized a pre-trained language model to learn text representations and calculated scores from the representations.

Table 5. Results of different models on ASAP dataset

	Quadratic Weighted Kappa Coefficient Values								
Models	1	2	3	4	5	6	7	8	Average
H1-H2	0.721	**0.812**	0.769	0.851	0.753	0.776	0.720	0.627	0.754
BLRR	0.761	0.606	0.621	0.742	0.784	0.775	0.730	0.617	0.705
CNN+LSTM	0.821	0.688	0.694	0.805	0.807	0.819	0.808	0.644	0.761
LSTM-CNN-att	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
RL1	0.766	0.659	0.688	0.778	0.805	0.791	0.760	0.545	0.724
SKIPFlOW	0.832	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
HISK+BOSWE	0.845	0.729	0.684	0.829	0.833	0.830	0.804	0.729	0.785
TSLF-ALL	0.852	0.736	0.731	0.801	0.823	0.792	0.762	0.684	0.773
R^2BERT	0.817	0.719	0.698	**0.845**	0.841	0.847	**0.839**	**0.744**	0.794
Our model	**0.884**	0.742	**0.807**	0.844	**0.867**	**0.879**	0.824	0.683	**0.816**

6. **Experimental results and analysis.** All models in Table 5 used the dataset provided by ASAP and evaluated the performance of the model by QWK. In the dataset, all subsets could be divided into expository and material essays based on genre type. The sentences in an explanatory essay have better articulation and coherence than in a material essay.

BLRR used Bayesian Ridge Regression, a machine learning model, to learn vocabulary-level features and use them to assess the quality of essays. CNN+LSTM used ordinary word embedding and extracted the average of all implicit states in the LSTM layer as the essay representation, which could not measure the quality of the essay comprehensively. LSTM-CNN-att treated essays as sentence-document hierarchies, fully considering the structural information between sentences and documents. SKIPFlOW used the parameters of the skipflow mechanism to act as auxiliary memory. Then, modelling the relationship between multiple locations allowed the model to learn representations of essays and approximate features of text coherence. LSTM-CNN-att and SKIPFlOW captured the explicit structure by modelling the relationship between the semantics of adjacent sentences in each essay so that they could score the explanatory essays more accurately (i.e., essay subsets 1, 2, 7, 8). RL1 had the lowest score among all models. Since it used extended LSTM to learn essay representation, it ignored sentence-level structural information. Therefore, for the scoring results of the material essay, the difference in the QWK was not significant for RL1 compared with LSTM-CNN-att and SKIPFlOW, which focused on the semantic relationship of adjacent sentences. However, the QWK of RL1 for the expository essay scoring results were lower than those of LSTM-CNN-att and SKIPFlOW. HISK+BOSWE achieved better results in essays' automatic scoring tasks than the previous deep learning models. The same showed the importance of shallow models in essay scoring. The two-stage model of TSLF-ALL proposed three scoring feature models, including semantic scoring, coherence scoring, and cue-related scoring. The experimental results showed a significant improvement compared with the models that only consider a single feature. R^2BERT combined the advantages of the RegressionOnly and RankingOnly models to fine-tune the Bert model and optimize it using multiple loss objectives. The self-attentiveness helped to master the concept of connectives and keywords in the essay. There was a significant improvement in the articulation between statements and the grasp of the theme of the essay. Thus the model achieved a good result in both subset and QWK.

Compared with other scoring models, the scoring model in this study achieves better results than other models in four subsets of essays. It shows that a good result could be achieved in different genres of essay categories. The QWK is improved by 0.111 compared to the machine learning model using only manual features alone (BLRR) and by 0.055 compared to the deep learning model only (CNN+LSTM). The advantage of the automatic scoring model based on multi-level vocabulary, sentence, and chapter adopted in this study is that it can represent the information of the whole essay at multiple levels. This scoring model also conforms to the scoring habits of the average marker, which provides a comprehensive analysis of the essay from multiple levels. Therefore, the proposed automatic essay scoring model has good generalization and optimal performance compared to the baseline model.

Table 6. Results of feature ablation on ASAP dataset

	Quadratic Weighted Kappa Coefficient Values								
Models	1	2	3	4	5	6	7	8	Average
vocabulary	0.842	**0.782**	0.737	0.753	0.812	0.756	0.749	0.659	0.761
sentence	0.856	0.641	0.683	0.784	0.756	0.816	0.772	0.615	0.740
chapter	0.765	0.648	0.668	0.713	0.744	0.727	0.703	0.498	0.683
vocabulary+sentence	**0.887**	0.711	0.776	0.842	0.847	0.863	0.803	0.607	0.792
vocabulary+chapter	0.873	0.758	0.790	0.809	0.835	0.846	0.797	0.620	0.791
sentence+chapter	0.859	0.673	0.769	0.827	0.836	0.842	0.796	**0.687**	0.786
Our model	0.884	0.742	**0.807**	**0.844**	**0.867**	**0.879**	**0.824**	0.683	**0.816**

In this study, ablation experiments are conducted to verify the effects of different levels of features on the performance of the scoring model. As can be seen from Table 6, in terms of the single-level feature model, the results of chapter-level feature based scoring are worse than those of the other two (vocabulary-level and sentence-level). From the score comparison between the single-level features, it can be seen that the vocabulary-level features have the largest proportion of improving the QWK, followed by the sentence-level features, and the chapter-level features are the smallest. It is difficult to distinguish between high and low-scoring essays simply based on chapter-level features, which can reflect a certain degree of information about the content of the essay but have a limited impact on the scoring results. However, from the scoring results of the combination of feature models at different levels, it can be seen that the combination of the sentence-level and the chapter-level with the vocabulary-level is higher than the score of the vocabulary-level alone. It can explain the importance of the sentence-level and chapter-level information of the essay in the scoring model. The combination of vocabulary, sentence, and chapter level features has the best results on most of the essay subsets, indicating that vocabulary-level features, sentence-level features, and chapter-level features all improve the model results to varying degrees.

Vocabulary-level features significantly impact the scoring model's performance, verifying that the combination of manual features and deep learning models can achieve a good result. Combining vocabulary-level features with chapter-level features improves the QWK by 0.108 over chapter-level features alone. Combining vocabulary-level features with sentence-level features improves the QWK by 0.052 over using sentence-level features alone. The experimental results show that vocabulary-level features can significantly improve the performance of the automatic scoring model for essays.

The impact of sentence-level features on the performance of the scoring model is second only to the effect of vocabulary-level features on the performance of the scoring model. The QWK improves by 0.103 after combining sentence-level features with chapter-level features than chapter-level features alone; the QWK improves by 0.031 after combining sentence-level features with vocabulary-level features than using vocabulary-level features alone. The experimental results show that sentence-level features can effectively improve the performance of the automatic scoring model for essays.

Chapter-level features have an impact on the performance of the scoring model. Compared with vocabulary-level features and sentence-level features, the QWK of using chapter-level features alone is the smallest, only 0.683. The combination of chapter-level features with vocabulary-level features improves the QWK by 0.03 over the vocabulary-level features alone. The combination of chapter-level features with sentence-level features improves the QWK by 0.046 over the sentence-level features alone. The experimental results show that the chapter-level features can enhance the performance of the essay scoring model, but the current chapter-level features have a limited impact on the performance of the AES model.

7. **Conclusions.** This study proposes an automatic scoring model for essays based on multi-level feature fusion. From a multi-level perspective, the CNN and BiLSTM are used to extract information at different vocabulary-level, sentence-level, and chapter-level, respectively, to improve the accuracy of essay scoring effectively. Experiments on the Kaggle ASAP competition dataset show that the proposed scoring model for essays outperforms other state-of-the-art models and verifies the advantage of the model in learning different types of features. In future research work, this study will start from the completeness of the AES model, combined with the essay content and other aspects to do further research.

Acknowledgment. This work is partially supported by Natural Science Foundation of Fujian Province under Award 2022J01133402, and Fujian Provincial Social Science Planning Project under Award
FJ2021BF050. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] E. B. Page, "Grading essays by computer: Progress report." in *Proceedings of the Invitational Conference on Testing Problems*, 1967.

[2] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *Journal of Computing in Higher Education*, vol. 33, no. 3, pp. 727–746, 2021.

[3] P. Wangkriangkri, C. Viboonlarp, A. T. Rutherford, and E. Chuangsuwanich, "A comparative study of pretrained language models for automated essay scoring with adversarial inputs," in *2020 IEEE Region 10 Conference (Tencon)*. IEEE, 2020, pp. 875–880.

[4] T. Zesch, M. Wojatzki, and D. Scholten-Akoun, "Task-independent features for automated essay grading," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 224–232.

[5] J. Lane and E. Lange, *Writing clearly: An editing guide.* Heinle & Heinle Boston, 1999.

[6] J. Shin and M. J. Gierl, "More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms," *Language Testing*, vol. 38, no. 2, pp. 247–272, 2021.

[7] S. Latifi and M. Gierl, "Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing," *Language Testing*, vol. 38, no. 1, pp. 62–85, 2021.

[8] X. Zhai, Y. Yin, J. W. Pellegrino, K. C. Haudek, and L. Shi, "Applying machine learning in science assessment: a systematic review," *Studies in Science Education*, vol. 56, no. 1, pp. 111–151, 2020.

[9] Z. Li, "Teachers in automated writing evaluation (awe) system-supported esl writing classes: Perception, implementation, and influence," *System*, vol. 99, pp. 102 505–102 519, 2021.

[10] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Machine learning towards intelligent systems: applications, challenges, and opportunities," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3299–3348, 2021.

[11] P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible domain adaptation for automated essay scoring using correlated linear regression," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 431–439.

[12] M. Mahana, M. Johns, and A. Apte, "Automated essay grading using machine learning," *Mach. Learn. Session, Stanford University*, vol. 5, 2012.

[13] H. Chen and B. He, "Automated essay scoring by maximizing human-machine agreement," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1741–1752.

[14] T. N. Tu, "An improving using mapreduce model in predicting learning ability of pupils based on bayes classification algorithm." *Journal of Information Hiding and Multimedia Signal Processing*, vol. 12, no. 3, pp. 140–151, 2021.

[15] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, "Application of quantum genetic optimization of lvq neural network in smart city traffic network prediction," *IEEE Access*, vol. 8, pp. 104 555–104 564, 2020.

[16] J. M.-T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C.-W. Lin, "A graph-based cnn-lstm stock price prediction algorithm with leading indicators," *Multimedia Systems*, pp. 1–20, 2021.

[17] M. Uto and M. Okano, "Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases," *IEEE Transactions on Learning Technologies*, vol. 14, no. 6, pp. 763–776, 2021.

[18] Y. Yang and J. Zhong, "Automated essay scoring via example-based learning," in *International Conference on Web Engineering.* Springer, 2021, pp. 201–208.

[19] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 715–725.

[20] F. Dong and Y. Zhang, "Automatic features for essay scoring–an empirical study," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1072–1077.

[21] Y. Tay, M. Phan, L. A. Tuan, and S. C. Hui, "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 5948–5955.

[22] Z. Ke and V. Ng, "Automated essay scoring: A survey of the state of the art." in *International Joint Conference on Artificial Intelligence*, vol. 19, 2019, pp. 6300–6308.

[23] J. Liu, Y. Xu, and Y. Zhu, "Automated essay scoring based on two-stage learning," *arXiv e-prints*, pp. arXiv–1901, 2019.

[24] T. Dasgupta, A. Naskar, L. Dey, and R. Saha, "Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring," in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 2018, pp. 93–102.

[25] K. Zupanc and Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems*, vol. 120, pp. 118–132, 2017.

[26] H. Nguyen and D. Litman, "Argument mining for improving the automated scoring of persuasive essays," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 5892–5899.

[27] M. Uto, Y. Xie, and M. Ueno, “Neural automated essay scoring incorporating handcrafted features,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6077–6088.

[28] Z. Shahbazi, Y.-C. Byun, and D. C. Lee, “Toward representing automatic knowledge discovery from social media contents based on document classification,” *International Journal of Advanced Science and Technology*, vol. 29, pp. 14 089–14 096, 2020.

[29] S.-M. Zhang, X. Su, X.-h. Jiang, M.-l. Chen, and T.-Y. Wu, “A traffic prediction method of bicycle-sharing based on long and short term memory network.” *Journal of Network Intelligence*, vol. 4, no. 2, pp. 17–29, 2019.

[30] J. M.-T. Wu, L. Sun, G. Srivastava, and J. C.-W. Lin, “A long short-term memory network stock price prediction with leading indicators,” *Big Data*, vol. 9, no. 5, pp. 343–357, 2021.

[31] J. M.-T. Wu, M.-E. Wu, P.-J. Hung, M. M. Hassan, and G. Fortino, “Convert index trading to option strategies via lstm architecture,” *Neural Computing and Applications*, pp. 1–18, 2020.

[32] S. Kumar, A. Damaraju, A. Kumar, S. Kumari, and C.-M. Chen, “Lstm network for transportation mode detection,” *Journal of Internet Technology*, vol. 22, no. 4, pp. 891–902, 2021.

[33] K. Taghipour and H. T. Ng, “A neural approach to automated essay scoring,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891.

[34] F. Dong, Y. Zhang, and J. Yang, “Attention-based recurrent convolutional neural network for automatic essay scoring,” in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017, pp. 153–162.

[35] Y. Wang, Z. Wei, Y. Zhou, and X.-J. Huang, “Automatic essay scoring incorporating rating schema via reinforcement learning,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 791–797.

[36] M. Cozma, A. M. Butnaru, and R. T. Ionescu, “Automated essay scoring with string kernels and word embeddings,” *arXiv e-prints*, pp. arXiv–1804, 2018.

[37] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1560–1569.