English Composition Topic and Emotion Classification Based on LDA-BERT Fusion and Dual GRU model

Lubing Shang
Anyang Vocational and Technical College, Anyang, Henan, 455000, China
E-mail: Lubing_Shang@outlook.com

Keywords: LDA model, BERT, GRU, theme recognition, sentiment polarity classification

Received: September 5,2025

With the development of globalization and artificial intelligence, the traditional manual correction of English compositions has problems, such as low efficiency and a lack of standardized evaluation. Therefore, there is an urgent need for automated analysis. The study aims to develop a precise model for identifying English essay themes and classifying emotional perspectives. Latent Dirichlet allocation is used to combine Probase prior knowledge and BERT fusion technology for topic modeling, and a Bi-GRU network with an attention mechanism is used for emotion classification. The experiment was carried out based on a dataset of 66,000 English compositions, including TOEFL, IELTS, and classroom compositions, and verified the effectiveness and superiority of the joint model in topic identification and emotion classification tasks. The experiment showed that the topic recognition model had a perplexity of 803, coherence of 0.831, and inter-class distance of 3.2 when there were 25 topics. When the threshold of the emotion classification model was 0.45, the accuracy reached 80.1% and the F1 value was 73.7%. The F1 value of 2,000 tests was 76.1%, which was better than the comparison model. Research has shown that this solution effectively solves problems such as insufficient semantic understanding in existing technologies, provides scientific support for intelligent evaluation systems, and promotes the intelligent development of educational evaluation.

Povzetek: Raziskava predstavi skupni model za samodejno prepoznavanje tem in čustvenih perspektiv v angleških esejih, ki z združitvijo LDA z BERT ter Bi-GRU z mehanizmom pozornosti dosega boljšo natančnost in semantično razumevanje pri inteligentnem vrednotenju pisnih nalog.

1 Introduction

With the deep integration of Artificial Intelligence (AI) and education, intelligent English composition evaluation systems have gradually become an important tool for assisting teaching [1]. However, the existing evaluation system has obvious shortcomings in the analysis from the emotional perspective. These systems cannot accurately grasp the emotional tendencies of students' compositions based on specific themes, resulting in a lack of depth and pertinence in the evaluation [2, 3]. Therefore, developing algorithms that can accurately identify the themes, emotions, and viewpoints of English composition sentences is of great significance for improving the quality of intelligent review and assisting teachers in comprehensively evaluating students' writing skills. Currently, advanced technologies in text sentiment analysis mainly rely on deep learning models, which perform well in general text sentiment classification tasks [4, 5]. Some researchers have attempted to incorporate attention mechanisms into the model to enhance its ability to capture key emotional information [6]. However, in the context of sentiment analysis in English compositions, most existing models overlook the guiding role of themes in emotional judgment, making it difficult to distinguish the emotional differences of the same expression under different themes. At the same time, there is a lack of refined modeling of the correlation between vocabulary and topic within sentences, resulting in a disconnect between sentiment analysis results and the core ideas of the composition [7]. Therefore, this study proposes a novel algorithm that integrates topic embedding and bidirectional neural networks. This algorithm deeply integrates the topic word embedding with the sentence word vector and utilizes Bidirectional Gated Recurrent Unit (Bi-GRU) to capture contextual information bidirectionally. It combines a topic-level attention mechanism to dynamically allocate vocabulary weights, ultimately achieving topic-based prediction of sentence sentiment probability distribution.

The innovation of the research lies in proposing the joint optimization of Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT). By preprocessing with Probase knowledge base and BERT deep semantic encoding, combined with neural encoding and decoding architecture, the topic recognition ability is enhanced. As an exception, a Bi-GRU sentiment classifier is synchronously designed, integrating topic attention mechanism to enhance contextual semantic modeling. This study adopts a dual model collaborative mechanism to achieve topic-oriented sentiment probability prediction through topic embedding and bidirectional network fusion, effectively breaking

through the limitations of traditional model semantic representation. This study aims to integrate the knowledge-enhanced topic model of Probase and BERT, combined with the topic-guided attention mechanism. It effectively solves the problems of shallow semantic understanding and separation of topic and emotion in traditional English composition analysis methods.

2 Related works

The current English composition topic recognition and Sentiment Polarity Classification (SPC) still faces problems such as blurred topic boundaries, misjudgment of emotional polarity, and insufficient understanding of context. Zhu proposed the Dirichlet process modeling+quantitative word frequency matrix method to efficiently utilize large amounts of text data in English curriculum development to extract key teaching themes. This method performed excellently in accuracy, recall, and precision [8]. Lin et al. proposed a probabilistic convolutional kernel implementation method based on memristors and a Bayesian forward training method to address the challenges faced by Bayesian Convolutional Neural Networks (CNNs) in hardware implementation. On the MNIST and CIFAR10 datasets, this method achieved accuracies of 98.67% and 87.81%, and achieved excellent out of distribution detection performance [9]. Cochran et al. proposed an identification method based on BERT and its variants to address the issue of causal relationship recognition and feedback automation in scientific explanation writing. The model could effectively identify the existence of causal relationships and was positively correlated with writing quality, but due to imbalanced data categories, there were still limitations in determining specific causal relationships, and further optimization was needed [10]. Cho et al. proposed a dual-scale BERT self-attention transformer CNN joint model to address the problem of insufficient modeling of overall and specific feature evaluation in automatic paper grading. This model combined dual-scale encoding with a multi-task learning framework. On the ASAP++ and TOEFL 11 datasets, this method improved the overall score by 2.0% and the multi-feature average performance by 3.6% compared to the single-task model [11]. Lagutina et al. proposed a vector model based on character, word, and syntactic features to automatically classify the language proficiency of short texts according to the Common European Framework of Reference for Languages. The F1 value of the support vector machine on the English corpus was 67%, while the basic version of the BERT computer-aided pattern evaluation and recognition joint method had the best performance in the BERT variant, with an F1 value of 69%. This indicated that the method had practical application potential [12].

Sharma et al. proposed a method that combines BERT pre-trained word embedding with Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) models to address the challenges of language understanding and feature expression in automatic essay scoring. The average quadratic weighted kappa coefficient of the SVR model reached 0.81, indicating its significant advantages in improving the accuracy and robustness of automatic scoring [13]. Rogers proposed an automatic optimization method based on a genetic algorithm to address the problem of difficult fair comparison between deep neural network architecture and hyperparameters in sentence classification tasks. The use of F1 value evolution models generally outperformed accuracy-based models, verifying the effectiveness in improving generalization ability and optimizing architecture selection [14]. Liu et al. proposed a deep learning model that combines residual networks and GRU to address the difficulties in feature extraction and limitations of traditional methods in the recognition of large and complex electrical objects. The average recognition rate was 95.56%, and it had good noise resistance [15]. Prakash et al. proposed a model based on CNN and GRU to address the difficulties in feature extraction and insufficient emotion discrimination in automatic speech emotion recognition. This method showed high recognition accuracy in three emotion state classification tasks, indicating its good robustness and generalization ability in speech emotion recognition applications [16]. Behrouzi et al. proposed a multimodal deep learning model based on GateGRU to address the complex classification of movie trailers and the limited performance of traditional models. This method significantly outperformed existing advanced models in classification performance, improving the accuracy and efficiency of multi-label movie type recognition [17].

In summary, the existing English composition topic recognition and SPC technology have problems such as shallow semantic understanding and incomplete feature extraction, making it difficult to accurately capture complex semantics and emotional tendencies. The LDA model and BERT topic recognition model have good semantic representation capabilities, while GRU utilizes Bi-GRU to enhance the capture of contextual information in text. Therefore, this study proposes an optimized LDA fusion BERT topic recognition model and an based on probase Bi-GRU SPC model. The collaboration between the two provides an efficient and accurate solution for English composition analysis.

Table 1: Comparison of research on existing methods

References	Research Focus	Methods Used	Dataset/Task	Key Indicators	Main Limitations / Differences from This Study
Text mining and topic modelling in English teaching: Extracting key themes and concepts for effective curriculum development.	Course Topic Extraction	Dirichlet Process Model	English Teaching Text	Accuracy, Recall, Precision	The association between topics and emotions is not considered, focusing on macro topic discovery.
Using BERT to identify causal	Causal	BERT and Its	Scientific Explanation	Causal	There is a problem of

structure in students' scientific explanations.	Relationship Recognition	Variants	Writing	Relationship Recognition Accuracy	category imbalance, and domain knowledge is not integrated for semantic enhancement.
Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading.	Automatic Essay Scoring	Dual-scale BERT-Self-Attention-CNN	ASAP++, TOEFL11	Overall Score Improvement by 2.0%	It focuses on overall scoring and does not deeply analyze the fine-grained association between topic structure and sentiment polarities.
Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model.	Text Difficulty Classification	SVM, BERT Variants	CEFR Standard English Corpus	F1 value (69%)	It focuses on language proficiency and does not involve the combination of topic modeling and sentiment analysis.
Modeling essay grading with pre-trained BERT features.	Automatic Essay Scoring	BERT + SVR/LSTM	Standard Essay Scoring Dataset	Weighted Kappa (0.81)	The model focuses on score prediction and lacks an explanation of the interactive mechanism between the internal topics and emotions of compositions.
A comparative analysis of deep neural network architectures for sentence classification using genetic algorithm.	Sentence Classification	Genetic Algorithm Optimized DNN	Standard Sentence Classification Task	F1 value	It focuses on the optimization of general architectures and does not design for the topic-emotion characteristics in the field of composition.

3 Methods and materials

3.1 Topic recognition model based on improved LDA and BERT feature fusion

With the acceleration of globalization and the development of AI technology, English writing ability has become a key skill for cross-cultural communication and academic research. In English composition evaluation, topic recognition is an important dimension for determining the relevance of content, and traditional LDA models are difficult to meet the evaluation needs of English compositions [18, 19]. Therefore, this study proposes an LDA topic recognition model that integrates BERT optimization. Before topic recognition, it is necessary to preprocess the essays to be submitted, as shown in Figure 1.

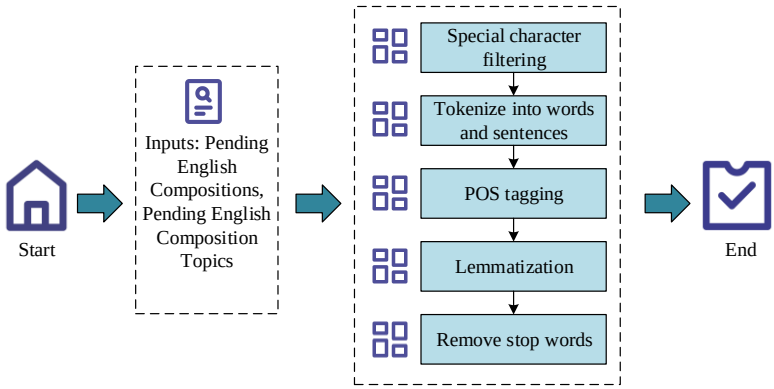


Figure 1: Preprocessing flowchart for essays awaiting approval

In Figure 1, the model needs to first read in the pending English essay and its title, and then filter out special characters using a special character set. Then, it uses the CoreNLP toolkit for word segmentation and sentence segmentation, and then performs part of speech tagging on the vocabulary after word segmentation and sentence segmentation. By combining the maximum entropy model with context to determine the part of speech, the different parts of speech annotated by vocabulary in different contexts are distinguished. Then, the vocabulary is restored to the dictionary prototype. Finally, stop words such as auxiliary verbs and prepositions are removed to obtain a collection of word roots that can represent the composition, laying the foundation for subsequent topic recognition and sentiment analysis [20, 21]. According to the root set obtained from the preprocessing module, the probability distribution of concept a in English composition is obtained, as shown in equation (1)

$$f(a_l|\mathbb{C}^{(t)}) = \frac{p(\mathbb{C}^{(t)}|a_l)\sum_{c_j^{(t)}\in a_l} n(a_l, c_j^{(t)})}{p(\mathbb{C}^{(t)})} \propto \sum_{c_j^{(t)}\in a_l} n(a_l, c_j^{(t)}) \prod_{j=1}^{N} p(c_j^{(t)}|a_l) \quad (1)$$

In equation (1), f is the posterior probability of the concept. C is the collection of pre-processed word roots in the composition. t is the t-th essay, a_l is the l-th concept, j is the root, and N is the total number of root sets. In the filtered set, the top three concepts in the probability distribution of the concept set of each English composition will be included in the global concept set of the corpus. Then, duplicate concepts in the set are removed, and concept clusters of corresponding themes are formed through concept clustering. The k-medoids algorithm is used for concept clustering, and cosine distance is used as the similarity measure. Through silhouette coefficient analysis, the optimal number of clusters is determined to be 8. Probase concepts are filtered based on word frequency and typicality scores, and the top 15% concepts with scores greater than 0.7 are retained to obtain concept space. Therefore, this study will integrate the optimized LDA and BERT pre-trained language model to construct an optimized LDA-fused BERT topic recognition model (LDA-BERT model, PL-BERT), as shown in Figure 2.

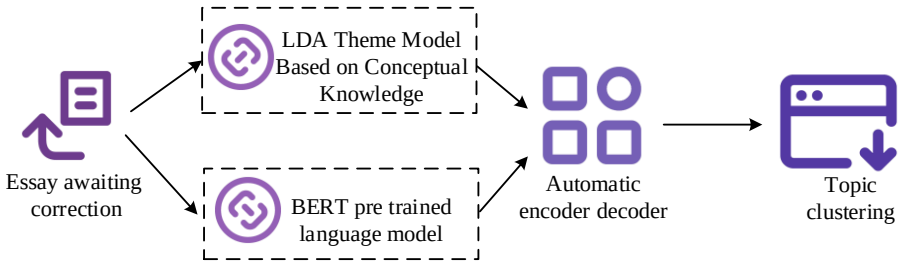


Figure 2: Topic recognition model based on optimized LDA fusion BERT (Source: https://iconpark.oceanengine.com)

According to Figure 2, the preprocessed English essay is first input into the LDA topic model based on the Probase concept knowledge base, and then a document concept set is generated using Probase. After filtering and clustering, it is used as an LDA prior parameter to obtain a topic vector that reflects global semantics. At the same time, the composition is processed by the BERT model, input in a special format, and output as a local semantic vector containing contextual syntactic information using the last layer of special classification tags. Then, the global topic vector and the local semantic vector are directly spliced, and the global topic distribution of LDA and the local context semantics of BERT are combined. The fusion vector L is shown in equation (2).

$$L = \text{gram}\{e_1, e_{2\ldots} e_B\} + \{f_1, f_{2\ldots} f_N\} \quad (2)$$

In equation (2), B is the number of topics, e is the essay topic vector, f is the essay semantic vector, and N is the number of semantics. The fusion vector is input into an automatic encoding decoding neural network topic model based on Gaussian distribution. The encoding layer generates topic vectors through multi-layer perceptrons, which are then transformed into topic probability distributions using softmax. The output latent topic vector R is shown in equation (3).

$$R = st\max(Av) \quad (3)$$

In equation (3), A is a linear transformation matrix, v is a vector with a Gaussian distribution, and st is a softmax function. The decoding layer reconstructs document semantics through potential topic vectors to generate documents, as shown in equation (4).

$$\log c\left(d_{r,i} \middle| \beta, b^r\right) = \log\left(b^r \cdot \beta\right) \quad (4)$$

In equation (4), d is location search, r is document, i is word, and β is topic vocabulary distribution matrix. b^r represents the topic distribution vector of the document. r and c are probabilities. The next step is to cluster each topic embedding representation using k-means clustering to obtain deep latent topic representations. Based on the above model structure, an optimized LDA fusion BERT based topic recognition algorithm can be obtained, as shown in Figure 3.

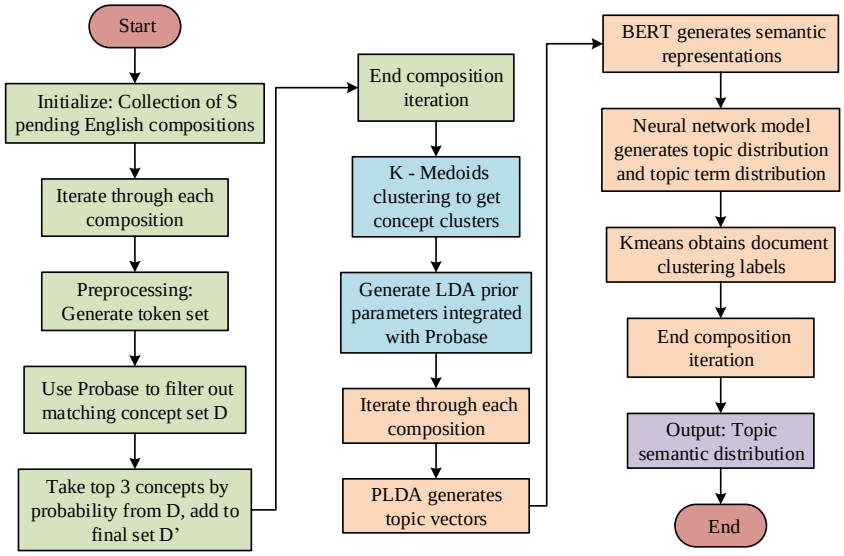


Figure 3: Theme recognition algorithm based on optimized LDA fusion BERT

In Figure 3, the BERT fusion algorithm first counts the number of articles, preprocesses each article, and then outputs a set of root words [22]. Next, the Probase concept knowledge base is utilized to generate a set of concepts that conform to the English composition corpus, and the top three probabilistic concepts are selected to form the set. The prior parameters of the Probase LDA Model (P-LDA) are obtained through clustering. Then, each essay uses the P-LDA model to obtain topic vectors and the BERT model to obtain contextual semantic vectors. The two are proportionally concatenated and input into a neural network topic model, which generates topic and term distributions through an automatic encoding-decoding structure. Finally, the topic tags are obtained through the k-means clustering algorithm, which enables in-depth mining of the hidden topic semantics of English compositions. The accuracy of topic identification and the completeness of semantic representation are significantly improved.

3.2 SPC model based on GRU network

Emotional perspective analysis, as an important dimension in English composition evaluation, directly affects the comprehensive evaluation of students' writing ability based on its accuracy. However, existing sentiment analysis models have significant shortcomings when dealing with complex texts such as English compositions [23, 24]. Bi-GRU can effectively extract long sequence text information. Therefore, this study proposes a Bi-GRU SPC model based on a topic-level attention mechanism, whose architecture is shown in Figure 4.

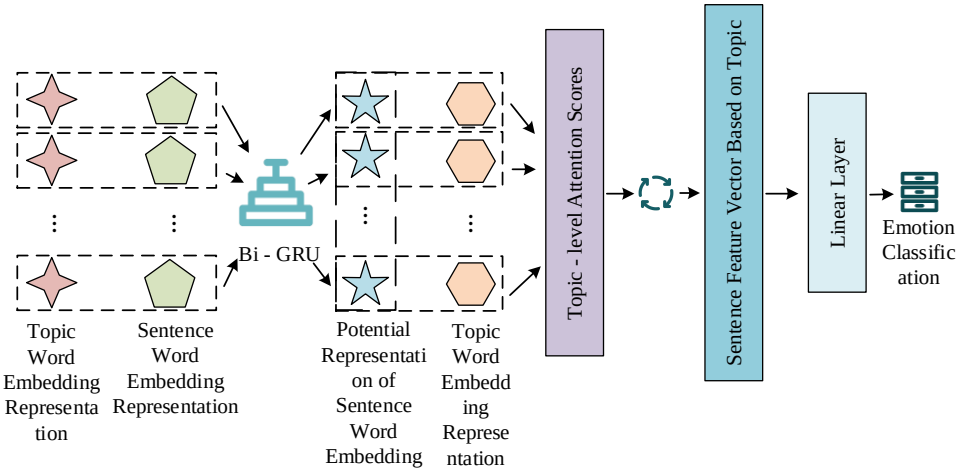


Figure 4: Schematic diagram of the optimized Bi-GRU SPC model

In Figure 4, the word embedding layer concatenates the sentence word embeddings generated by the Word2Vec model with the latent topic vectors of the pending essay to form the model input. The next step is to process the input sequence using Bi-GRU to obtain the contextual semantic latent representation of each term. The topic-level attention mechanism layer calculates the attention scores of each vocabulary in the sentence based

on the topic embedding vector, dynamically adjusts the weights, and weights the sum to focus on the key features that express the topic's emotions. Among them, the theme embedding vector is shown in equation (5).

$$T_o = \tanh(q_s y_o + p_s) \tag{5}$$

In equation (5), q_s is a $d_e \times d\_h$ matrix. y_o and q_s are all of d dimensions. T_o is the output vector, q_s is the weight matrix, y_o is the input hidden state vector, and p_s is the bias vector. The calculation of the attention score for each vocabulary on the topic is shown in equation (6).

$$k_o = \frac{\exp(u_o \cdot h)}{\sum_{j=1}^{m} \exp(u_j \cdot h)}, \sum_{o=1}^{m} k_o = 1 \tag{6}$$

In equation (6), u is a d_h-dimensional vector, h is of d_h dimension, and k outputs a 1-dimensional value. k is the attention weight, u_o represents the intermediate semantic vector of the o -th word, and ex is the transformation function. j represents the index of the j -th word, h is the queryable vector, and m is the total number of elements involved in attention calculation. The latent semantics of the sentence are shown in equation (7).

$$d = \sum_{i=1}^{s} k_i y_o \tag{7}$$

In equation (7), d is of d_h dimension. d is the final eigenvector obtained by weighted summation, as shown in equation (8).

$$z = soft\max(sigmoid(Y_s, s)) \tag{8}$$

In equation (8), Y is a $d\_out \times d\_h$ matrix. s is the d_h dimension, and z is the d_out dimension. The output dimension of the activation function is the same as its input dimension, and no Dropout layer is used. z is the final probability of outputting sentiment classification, s is the latent semantics of the sentence, and Y is the change matrix. This architecture provides key criteria for evaluating the quality of essay themes and viewpoints through inter layer collaboration. Based on the above architecture, this study has improved the topic-based sentiment classification algorithm, as shown in Figure 5.

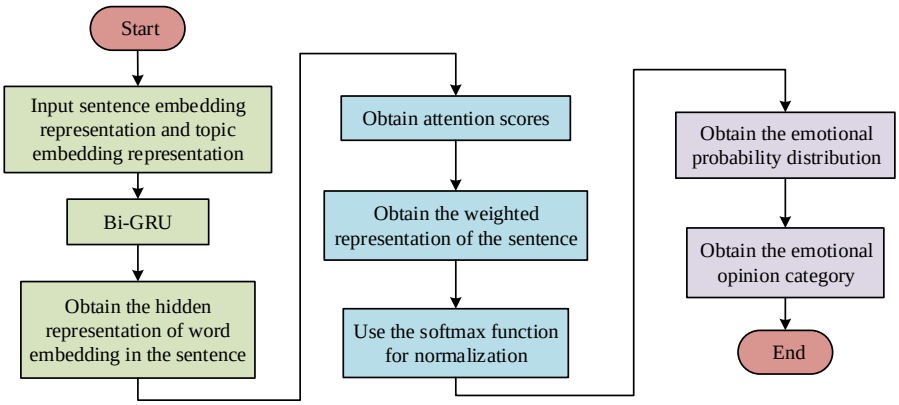


Figure 5: Flowchart of theme-based emotion classification algorithm

In Figure 5, the algorithm first inputs the word embedding representation set and the topic word embedding representation of the English composition sentence to be approved. Next, the topic word embedding and sentence word embedding are concatenated and input together into the Bi-GRU to obtain the hidden representation of the sentence word embedding. Subsequently, the topic-level attention mechanism is used to process the hidden representation set, and the topic vector of the sentence is used as the "query" benchmark. The correlation between the hidden state of each word in the sentence and the topic vector is calculated to obtain the attention score. Based on the attention score, the hidden representation set is weighted and summed to form a sentence representation that can dynamically focus on topic-related emotional words. Then, the weighted representation is processed through linear layers and softmax functions to achieve the probability distribution of emotions related to the topic. Finally, the tagging process ends and returns the sentiment probability distribution, which is used to present the results of sentiment analysis based on the topic of the sentence. To analyze the effectiveness of the themes and sentiment polarities obtained from relevant models, this study designs an English composition theme viewpoint quality evaluation method. This method is used to score the themes and sentiment polarities of the composition, as shown in Figure 6.

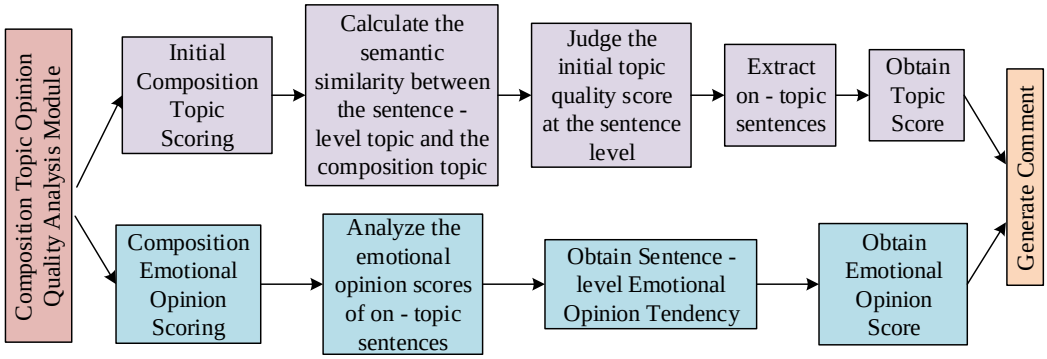


Figure 6: Quality evaluation method for English composition theme viewpoints

In Figure 6, in the topic relevance path, step 1 is to extract the semantic vectors of the sentence and the essay topic through the BERT model, and use the cosine similarity algorithm to calculate the semantic matching degree between the two. The scoring mechanism is constructed based on this similarity to quantify the sentence-level topic quality score, achieve accurate recognition of the target text, select the target sentence, and ultimately obtain the topic score. The sentiment polarity path first analyzes the sentiment polarity scores of relevant sentences, delves into the emotional tendencies of related content, and then obtains sentence-level sentiment polarity scores. After the theme rating and sentiment polarity rating are determined, the system will generate comments and integrate these indicators to produce a comprehensive and personalized evaluation. This method constructs a dual-focus evaluation system, ensuring a close fit between the composition and the theme, while taking into account the richness and relevance of the emotional perspective. This provides a comprehensive evaluation perspective for the quality of English composition topic perspectives.

4 Results

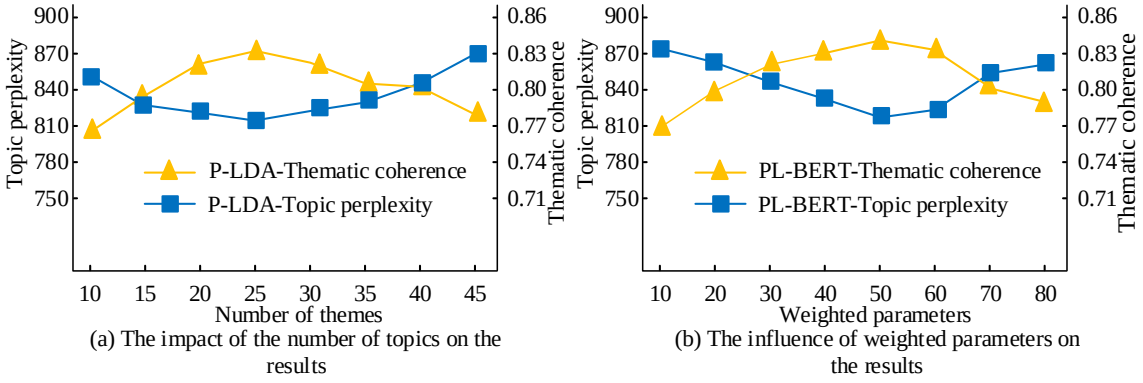
4.1 Topic recognition model test results

To conduct experimental testing of the designed model, the study uses 40,000 EFCAMDAT, 16,000 TOEFL11 Corpus, and 10,000 IELTS Essay datasets, totaling 66,000 English essays. The study divides 66,000 essays into training sets and test sets at a ratio of 8:2. The model is trained using the Adam optimizer, the learning rate is set to 5e-5, the batch size is 32, and a total of 50 rounds are trained. An early stopping method is used to prevent overfitting. This study also builds a simulation testing platform to test the performance of the model. The platform is tested with SpaCy 3.5.0 and processes with Pandas 2.1.0. The detailed information is shown in Table 2.

Table 2: Detailed information of the testing platform

Hardware Information		Software information	
Name	Type	Name	Type
CPU	AMD Threadripper 3990X	Text preprocessing tool	NLTK 3.8.1
GPU	NVIDIA A100	Natural Language Processing Library	SpaCy 3.5.0
Storage device	1TB NVMe SSD+4TB SATA SSD	Data processing tools	Pandas 2.1.0
Dynamic Calculation Chart Tool	PyTorch 2.0	Virtual Environment Tools	Anaconda 2023.07
Static Calculation Chart Tool	Pandas 2.1.0	Development tool	PyCharm 2023.3

All comparative experiments are independently repeated five times under identical conditions. The final reported results are averaged, and key indicators are tested with two-sided t-tests. To test the performance of PL-BERT in topic recognition of English compositions, this study uses topic perplexity and topic coherence as judgment indicators. The experimental comparison between the number of topics in P-LDA and the vector weighting parameters in PL-BERT is shown in Figure 7.

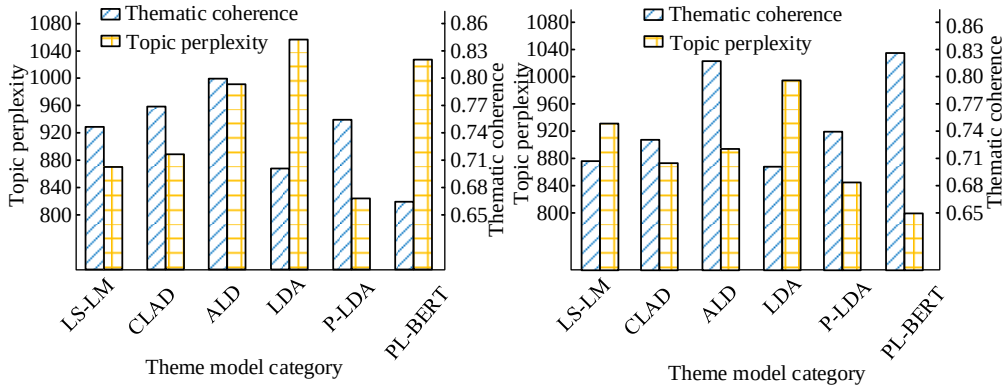


(a) The impact of the number of topics on the results

(b) The influence of weighted parameters on the results

Figure 7: The impact of the number of themes and weighting parameters on the model

In Figure 7 (a), as the number of topics gradually increases from 10, the topic perplexity of P-LDA gradually decreases from 857. When the number of topics is 25, the topic perplexity drops to the lowest, which is 803. As the number of topics increases, the level of thematic confusion also gradually increases. The thematic coherence reaches its peak when the number of themes is 25, with a thematic coherence of 0.831. Therefore, under the experimental data, when the number of topics is 25, the model has the best clustering effect on English essay topics. In Figure 7 (b), when the weighting parameter is 50, the topic perplexity is the lowest, at 821. At this point, the coherence of the theme reaches its highest value of 0.864, indicating that when the weighting parameter is 50, PL-BERT can achieve optimal performance. When the number of topics deviates from 25 or the weighted parameters deviate from 50, both indicators fluctuate, which also verifies the rationality of the parameter settings. The metric values at parameter 25 and parameter 50 show statistically significant differences from adjacent parameters (p<0.05). The proposed PL-BERT is compared with Large-Scale Language Models (LS-LM) [18], Computer Language-Aided models (CLAD) [19], Automatic Language Detection (ALD) [20], LDA, and P-LDA models, as shown in Figure 8.



(a) Evaluation results of different models on daily classroom composition

(b) Evaluation results of different models on TOEFL and IELTS essay writingdata

Figure 8: Differences in model influence under different datasets

In Figure 8 (a), in the daily classroom composition dataset, the theme perplexity of PL-BERT is about 821, with a theme coherence of 0.823, while the perplexity of P-LDA is about 824, with a coherence of about 0.754. In the experiment, the topic perplexity of LS-LM, CLAD, and ALD models is higher than that of PL-BERT, and the topic coherence is lower than that of PL-BERT. In Figure 8 (b), in the TOEFL and IELTS essay datasets, PL-BERT has a perplexity level of approximately 803 and a coherence level of approximately 0.831. Its performance in testing topic perplexity and coherence is still superior to other models. Whether in the daily classroom composition dataset or the TOEFL/IELTS composition dataset, PL-BERT exhibits good performance advantages in different language usage contexts. This also validates the effectiveness and robustness of the model fusion strategy on different datasets. The PL-BERT model outperforms all baseline models by a statistically significant margin (p<0.01) on both datasets. To verify the theme clustering effect of PL-BERT, this study uses TOEFL/IELTS essays as experimental data. The experiment uses inter-class distance and intra-class compactness as evaluation indicators, selects four similar topics (dividing into 8 vocabulary words), and conducts visual clustering analysis based on 2D UMAP technology [21]. The traditional LDA model is used as the comparison benchmark, as shown in Figure 9.

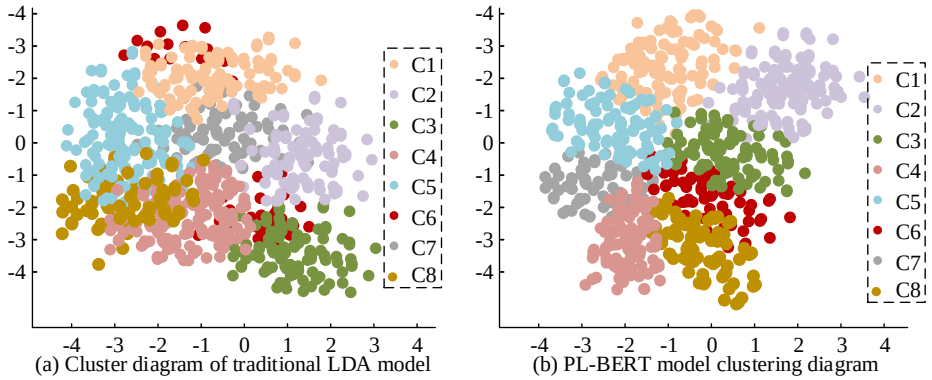


Figure 9: Clustering effect of different models (Image source: self drawn by the author)

In Figure 9 (a), the clustering points of LDA are loosely distributed, with an inter-class distance of only 1.8 and an intra-class compactness of 0.68. The silhouette coefficient of the traditional LDA model is only 0.28. Different topic clusters overlap, and some topic clusters exhibit obvious separation. Among the distribution characteristics of topic clusters, LDA has obvious shortcomings in classifying topic clusters. In Figure 9 (b), the PL-BERT clustering points exhibit clear inter-class separation and high intra-class clustering. The distance between topic clusters under PL-BERT is 3.2, and the compactness of samples within the same topic is 0.45. The contour coefficient of the PL-BERT model has been increased to 0.62. Compared with the inter-class distance index and intra-class compactness of the two models, PL-BERT has a higher inter-class distance, clear differentiation, lower intra-class compactness, and clear division of topic samples. This validates the advantages of the model in semantic space mapping and topic clustering. The PL-BERT model outperforms the LDA model significantly in both inter-class distance and intra-class compactness (p<0.001).

4.2 Opinion classification test results

The study conducts experiments using data from the Common European Language Reference Framework (CEFR) at levels B1 to C1 in the test set. To verify the experimental effectiveness of the emotional perspective analysis model, this study tests and evaluates the emotional perspective attitude data extracted from English compositions on various topics. This study uses precision, recall, and F1 value as evaluation indicators to conduct comparative experiments between the optimized Bi-GRU SPC model and the topic-based sentiment classification model, as shown in Figure 10.

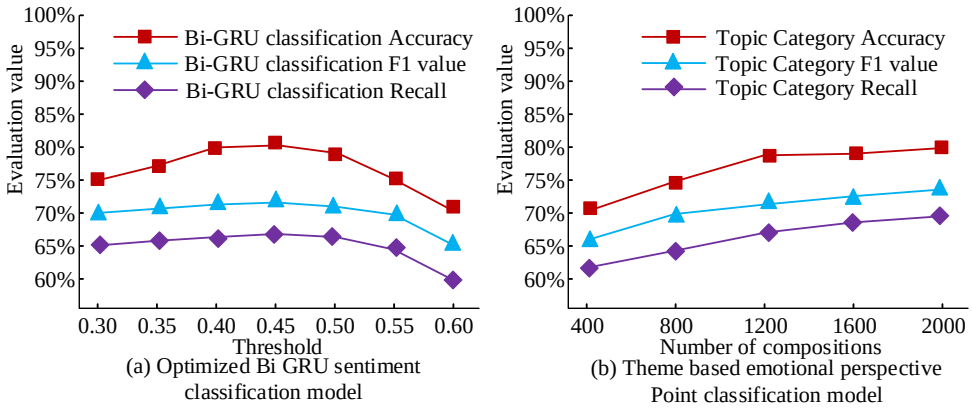


Figure 10: Comparison of model index testing under different conditions

In Figure 10 (a), the optimal accuracy, recall, and F1 value of the optimized Bi-GRU SPC model can be tested by adjusting the threshold. As the threshold increases, each indicator shows a trend of first increasing and then decreasing. When the threshold reaches 0.45 (p<0.05), the accuracy, recall, and F1 value of the model reach 80.1%, 68.2%, and 73.7%, demonstrating the best performance in predicting positive cases and capturing positive cases in experiments. In Figure 10 (b), as the number of essays increases from 400 to 2,000, the index indicators of optimal accuracy, recall, and F1 value all show a gradually increasing trend. However, within the range of 400-1,200 essays, the growth rate of each indicator value is significantly faster than that of the following 1,200-2,000. However, this experiment has not yet tested the limit on the number of articles with peak values for each indicator, demonstrating that the model has good compatibility with data capacity and stable robustness. To verify the performance advantages of the theme SPC model, this study selects the Based on the Recurrent Neural Network

Classification Model (BCNNC) [22], Evidence is Stacked on the GUR Neural Network (E-GUR) [23], and Neural Network Fusion Model (NNF) [24] as baselines. This experiment compares accuracy, precision, recall, and F1 value indicators. The test results are shown in Figure 11.

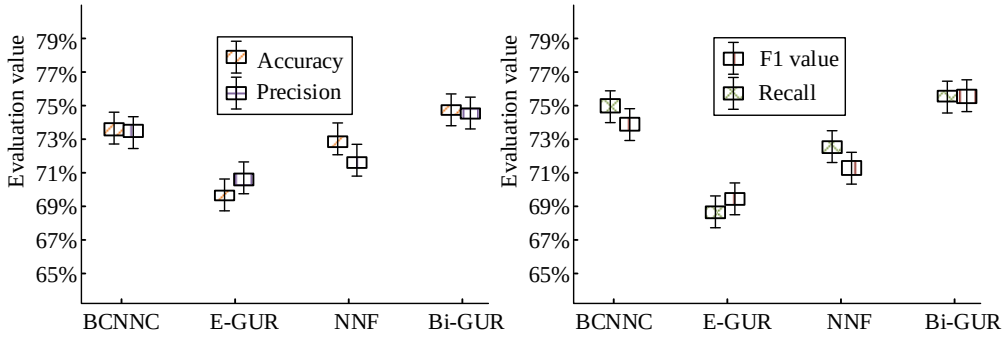


Figure 11: Comparison of indicators for different models (Image source: self drawn by the author)

In Figure 11 (a), among all the models involved in the experiment, the precision and accuracy of Bi-GUR are 75.9% and 75.4%, slightly higher than the 73.5% and 73.8% of BCNNC. The precision and accuracy of E-GUR are the lowest, at 70.8% and 70.1%. Bi-GUR, which has stronger comprehensive performance value and environmental adaptability, performs more outstandingly. In Figure 11 (b), the recall rate and F1 value of Bi-GUR are 76.3% and 76.1%. The E-GUR with the worst performance in this part of the experiment is 69.8% and 70.6%, which are 5.5% lower than the corresponding indicators of Bi-GUR. Further comparison with other models shows that Bi-GUR has a recall index 4.2% higher than the theme level NNF, and an F1 value 0.8% higher than BCNNC. Overall, Bi-GUR outperforms most comparison models in all indicators and demonstrates significant advantages in English composition SPC tasks. Especially in terms of recall rate and F1 value, it demonstrates the effectiveness of capturing contextual semantic dependencies through Bi-GRU and enhancing emotional feature extraction through topic-level attention mechanism (p<0.05). Bi-GUR can more comprehensively identify sentiment polarity information in essays.

5 Discussion

The LDA-BERT and Bi-GRU integrated model demonstrates exceptional performance in the tasks of topic identification and sentiment classification of English essays. Its core advantage stems from the synergistic innovation in its architecture. This study achieves deep semantic association by constructing a dual-model framework. In terms of topic identification, it innovatively incorporates the Probase concept knowledge base as a semantic prior for LDA, effectively addressing the data sparsity issue of traditional LDA models in short text processing. Simultaneously, deep integration with the BERT model leverages its powerful contextual semantic understanding capability to compensate for LDA's shortcomings in handling polysemous words and complex syntactic structures. The latent topic vectors generated by the neural encoder after the fusion of these two components possess both global topic distribution and local semantic details, significantly enhancing topic identification performance. For sentiment classification, a Bi-GRU model based on a topic-level attention mechanism is designed. By using topic embedding vectors extracted from PL-BERT as queries for the attention mechanism, the model dynamically focuses on emotional vocabulary highly relevant to the core topic. Compared to existing advanced methods, the fundamental breakthrough of this research lies in achieving a deep association between topics and sentiments. The Bi-GRU model performs well in terms of overall article scoring, but it ignores fine-grained topic-emotion correlations, and the BERT+SVR/LSTM combination lacks explicit modeling of the topic-emotion interaction mechanism within the article. In contrast, this study achieves a performance breakthrough through the following mechanism. The introduction of Probase prior knowledge guides the model to focus on semantically clear concept clusters, significantly increasing topic coherence to 0.831. The deep semantic fusion of BERT accurately disambiguates polysemous words and complex sentence structures, reducing model perplexity to 803. The topic-level attention mechanism prioritizes emotional expressions highly relevant to the core topic, significantly improving recall while maintaining a precision of 80.1%, ultimately achieving an F1 value of 76.1%. However, the study still has limitations, including insufficient cross-linguistic generalization capability, limited robustness to noisy input, and difficulties in understanding complex rhetorical devices such as metaphors and irony. Future work should focus on lightweight design, cross-language transfer learning, multi-modal fusion, dynamic update mechanism, etc., to enhance the applicability and scalability of the model under different load conditions. This promotes the development of intelligent English composition evaluation in a more accurate and efficient direction.

6 Conclusion

In the context of intelligent education, traditional English composition evaluation urgently needs to innovate with automated analysis technology due to its efficiency and objectivity deficiencies. This study focuses on topic recognition and emotion classification, and constructs a dual model framework of "optimized LDA fusion BERT" and "optimized Bi-GRU". The topic recognition model achieves deep topic extraction through Probase concept screening, BERT semantic vector fusion with LDA prior parameters, and neural network encoding and decoding. The sentiment classification model utilizes Bi-GRU and a topic-level attention mechanism to dynamically capture text sentiment dependencies. Experimental verification showed that the topic model performed outstandingly on the TOEFL/IELTS dataset. When the number of topics was 25, the perplexity was 803, the coherence was 0.831, the inter-cluster distance was 3.2, and the intra-class compactness was 0.45, which was better than the traditional LDA model. The advantages of semantic space mapping and topic clustering were significant. At a threshold of 0.45, the emotion model had an accuracy of 80.1% and an F1 value of 73.7%. The F1 value in the 2,000-essay test increased to 76.1%, which was 4.2% higher than the recall rate of BB-TWTR. The robustness was stable under different data capacities. The model designed in this study breaks through the limitations of the semantic understanding of traditional methods by strengthening semantic representation and topic association. However, its architecture is designed for English and lacks cross-language generalization capabilities. It has limited robustness to noisy input containing spelling errors and non-standard expressions. Moreover, it still has obvious deficiencies in understanding complex rhetorical techniques such as metaphor and irony. Future work can enrich the emotional dimension with multi-modal features by introducing cross-language transfer learning, and integrate a dynamic word vector update mechanism that adapts to language evolution. It can also expand application scenarios and promote the development of intelligent English composition evaluation in a more accurate and diversified direction.

References

[1] Istiqomah A A, Sari C A, Susanto A, et al. Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50. Journal of Applied Informatics and Computing, 2024, 8(2): 257-264. DOI: 10.30871/jaic.v8i2.8329.

[2] Wadawadagi R, Tiwari S, pagi V. Polarity-aware deep attention network for aspect-based sentiment analysis. Progress in Artificial Intelligence, 2025, 14(1): 33-48. DOI: 10.1007/s13748-024-00352-x.

[3] Kristina K, Suarjaya I M A D, Wiranatha A A K A C. Stock Sentiment Prediction of LQ-45 Based on News Articles Using LSTM. Journal of Applied Informatics and Computing, 2025, 9(4): 1154-1162. DOI: 10.30871/jaic.v9i5.10157.

[4] Jones C, Roulet V, Harchaoui Z. Revisiting Convolutional Neural Networks from the Viewpoint of Kernel-Based Methods. Journal of Computational and Graphical Statistics, 2023, 32(4): 1237-1247. DOI:10.1080/10618600.2022.2163649.

[5] Xue M L, Wang D Y. Cross-modal sentiment analysis on social media using improved nonverbal representation learning and GHRNN fusion. Informatica, 2025, 49: 35-50. DOI: https://doi.org/10.31449/inf.v49i34.9130.

[6] Wang S, Li F. Deep Reinforcement Learning with Convolutional Neural Networks for Optimizing Supply Chain Inventory Management. Informatica, 2025, 49: 201-222. DOI: https://doi.org/10.31449/inf.v49i26.8396.

[7] Shi J, Li Z, Lai W, LI F, Shi R, Feng Y, et al. Two end-to-end quantum-inspired deep neural networks for text classification. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 4335-4345. DOI:10.1109/TKDE.2021.3130598.

[8] Zhu J. Text mining and topic modelling in English teaching: Extracting key themes and concepts for effective curriculum development. Journal of Computational Methods in Sciences and Engineering, 2025, 25(2): 1210-1222. DOI: 10.1177/14727978241298468.

[9] Lin Y, Zhang Q, Gao B, Tang J, Zhao H, Qin Q, et al. High-Efficient Memristor-Based Bayesian Convolutional Neural Networks for Out-of-Distribution Detection by Uncertainty Estimation. IEEE Transactions on Electron Devices, 2025, 72(1): 206-214.DOI:10.1109/TED.2024.3497917.

[10] Cochran K, Cohn C, Hastings P,Tomuro N, Hughes S . Using BERT to identify causal structure in students' scientific explanations. International Journal of Artificial Intelligence in Education, 2024, 34(3): 1248-1286. DOI: 10.1007/s40593-023-00373-y.

[11] Cho M, Huang J X, Kwon O W. Dual‐scale BERT using multi‐trait representations for holistic and trait‐specific essay grading. ETRI Journal, 2024, 46(1): 82-95. DOI:10.4218/etrij.2023-0324.

[12] Lagutina N S, Lagutina K V, Brederman A M, Kasatkina N N. Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model. Automatic Control and Computer Sciences, 2024, 58(7): 869-878. DOI: 10.3103/S0146411624700329.

[13] Sharma A, Jayagopi D B. Modeling essay grading with pre-trained BERT features. Applied Intelligence, 2024, 54(6): 4979-4993. DOI: 10.1007/s10489-024-05410-4.

[14] Rogers B, Noman N, Chalup S, Moscato P. A comparative analysis of deep neural network architectures for sentence classification using genetic algorithm. Evolutionary Intelligence, 2024, 17(3): 1933-1952. DOI: 10.1007/s12065-023-00874-8.

[15] Liu S, Xing L, Hao X, Gong S, Xu Q, Qi W. Electrically Large Complex Objects Recognition Based on Gated Recurrent Residual Network (GRRNet). IEEE Open Journal of Antennas and

Propagation, 2025, 6(2): 365-371. DOI:10.1109/OJAP.2024.3516835.
[16] Prakash P R, Anuradha D, Iqbal J, Galety M. G., Singh R., Neelakandan S. A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification. Journal of Control and Decision, 2023, 10(1): 54-63. DOI:10.1080/23307706.2022.2085198.
[17] Behrouzi T, Toosi R, Akhaee M A. Multimodal movie genre classification using recurrent neural network. Multimedia Tools and Applications, 2023, 82(4): 5763-5784. DOI: 10.1007/s11042-022-13418-6.
[18] Yavuz F, Çelik Ö, Yavaş Çelik G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 2025, 56(1): 150-166. DOI:10.1111/bjet.13494.
[19] Tariq U. Nexus of essay writing and computer-assisted language learning (CALL) in English language classroom. Interactive Technology and Smart Education, 2025, 22(1): 103-133. DOI: 10.1108/ITSE-12-2023-0246.
[20] Chan K K Y, Bond T, Yan Z. Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. Language Testing, 2023, 40(1): 61-85. DOI: 10.1177/02655322221076025.
[21] Becht E, McInnes L, Healy J, Dutertre C, Kwok I, Ng L, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology, 2019, 37(1): 38-44.DOI:10.1038/nbt.4314.
[22] Jeong S, Chung W, Shin S, Kim D, Kim J, Byun G, et al. Development of Dolphin Click Signal Classification Algorithm Based on Recurrent Neural Network for Marine Environment Monitoring. Geophysics and Geophysical Exploration, 2023, 26(3): 126-137. DOI:10.7582/GGE.2023.26.3.126.
[23] Zhou H, Chen W, Liu J, Cheng L, Xia M. Trustworthy and intelligent fault diagnosis with effective denoising and evidential stacked GRU neural network. Journal of Intelligent Manufacturing, 2024, 35(7): 3523-3542. DOI: 10.1007/s10845-023-02221-1.
[24] Rabbani M H R, Islam S M R. Deep learning networks based decision fusion model of EEG and fNIRS for classification of cognitive tasks. Cognitive Neurodynamics, 2024, 18(4): 1489-1506. DOI: 10.1007/s11571-023-09986-4.