Applying the BERT Model Combined with Multi-Task Learning to Improve the Accuracy of Academic Automatic Scoring in Higher Education Paper Writing

Danni Xie*

Hunan University, Changsha, 413100, China

*x_dn2025@163.com

Abstract. The current automatic scoring of academic content in higher education paper writing has problems such as insufficient accuracy and inconsistent scoring standards. To address these challenges, this paper proposes an innovative method that combines the BERT model and multi-task learning. Specifically, this paper uses the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model as the basis for text representation, making full use of its powerful contextual semantic understanding ability to accurately analyze the academic content of papers. Next, in view of the multiple dimensions involved in academic scoring, this paper designs a multi-task learning framework that allows the model to share knowledge between different tasks, thereby improving the performance of the model in scoring in each dimension. In order to further enhance the robustness of the model, this paper also adds a data augmentation strategy during the training process, using multiple manually annotated scoring samples for joint training. The experimental results show that the accuracy of the model in terms of logic scoring is 84.1%, and in terms of language quality, it is 87.5%. The experiment also verifies that the method has strong generalization ability in cross-domain adaptability, especially when dealing with papers from different disciplines, it is more stable and accurate. In summary, the method proposed in this paper provides a more accurate and consistent solution for academic automatic scoring in higher education paper writing.

Keywords: BERT Model; Academic Scoring; Contextual Semantic Understanding; Educational Artificial Intelligence; Multi-Task Learning

1 Introduction

As the scale of higher education continues to expand, the number of student papers has grown rapidly. Traditional manual scoring methods face significant challenges in terms of labor costs, scoring efficiency, and subjective control. In recent years, automatic scoring technology has gradually emerged in educational scenarios. However, existing methods mostly focus on surface language features and have difficulty in deeply understanding the semantic logic and structural norms in academic writing, resulting in a large room for improvement in scoring accuracy and consistency. Therefore, the study

© The Author(s) 2025
A. Y. M. A. Islam et al. (eds.), *Proceedings of the 2025 4th International Conference on Science Education and Art Appreciation (SEAA 2025)*, Advances in Social Science, Education and Humanities Research 952,
https://doi.org/10.2991/978-2-38476-452-5_14

of automated models that can effectively identify the deep academic features of texts and have multi-dimensional scoring capabilities has important practical needs and theoretical value.

The full text is structured as follows: the first part introduces the research background and related work, clarifies the problem positioning and research objectives; the second part describes in detail the overall architecture and key technical design of the proposed model; the third part shows the experimental settings and results comparison, verifying the model's performance in accuracy, consistency and cross-domain adaptability; the last part summarizes the full text work and proposes future research directions that can be expanded.

2 Related Work

The research on automatic scoring systems for academic papers has a long history, and many scholars have proposed different scoring methods. For example, Messer et al. [1] conducted a systematic literature review of 121 research papers on automatic scoring and feedback tools for programming education from 2017 to 2021, and classified them according to evaluation skills, methods, language paradigms, degree of automation, and evaluation techniques. Ramesh and Sanampudi [2] conducted a systematic literature review of automatic paper scoring systems, studied artificial intelligence and machine learning technologies used to evaluate automatic paper scoring, and analyzed the limitations and development trends of current research, pointing out that paper evaluation is not based on the relevance and coherence of content. Automatic grading and feedback tools have become a key component of the rapid growth of online learning. Hahn et al. [3] analyzed 125 research results on automatic grading and feedback as learning tools published in journals and proceedings between 2016 and 2020. Kooli and Yusuf [4] highlighted the advantages and potential risks of artificial intelligence in education, especially the possibility of exacerbating educational inequality, and pointed out the importance of striking a balance between artificial intelligence and human teaching to achieve optimal educational outcomes. Combéfis [5] reviewed the automatic code evaluation systems in recent years and pointed out that live courses and online courses developed for distance learning require automatic grading in order to effectively scale and manage a large number of learners with limited teaching staff. Wilson et al. [6] investigated the implementation of an automated writing evaluation software called MI Write in elementary schools. The results provided inspiration for other school districts to implement AWE (Automated Writing Evaluation) and design AWE systems.

3 Method

3.1 Building a BERT-Based Paper Semantic Representation Model

As a pre-trained deep learning model, the BERT model can capture contextual information at the word, sentence, and even paragraph level by pre-training on a large-scale

text corpus. Unlike traditional unidirectional language models, BERT uses a bidirectional encoder structure that can simultaneously consider contextual information, thereby effectively understanding the meaning of words in different contexts. The core of the BERT model is the self-attention mechanism based on the Transformer encoder, and its calculation formula is as follows (1):

$$Z = \text{SelfAttention}(Q, K, V) \tag{1}$$

Q, K, and V represent query, key, and value, respectively. This feature enables BERT to better capture the implicit grammatical structure and logical relationships in academic texts, especially when analyzing the rigor of arguments, argumentation process, and structured information in academic papers, showing significant advantages.

Specifically, the input of the BERT model is a text sequence, which is converted into a high-dimensional vector through word embedding. BERT converts each word of the input text into a vector representation with contextual semantics through its multi-layer Transformer encoder. After BERT processing, the representation of each word not only contains the semantic information of the word itself but also the contextual information of the word in the sentence where it is located, which enables the model to capture more complex semantic relationships. This process is shown in Figure 1:

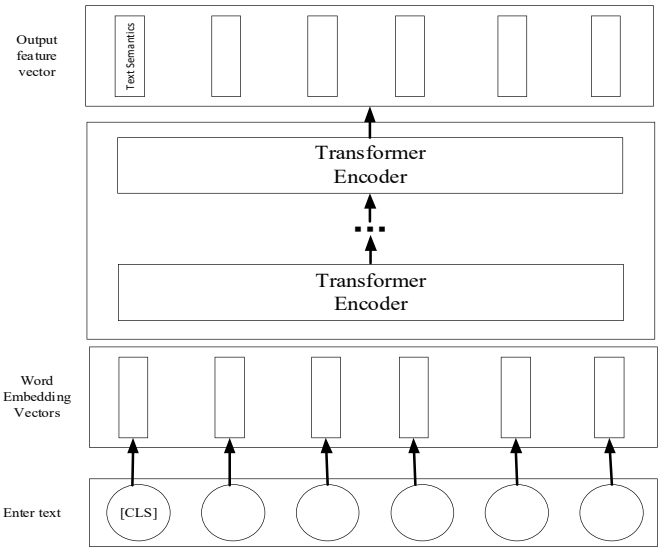

Fig. 1. BERT model structure diagram

In the experiments in this paper, a pre-trained BERT model is used and fine-tuned to adapt to the characteristics of higher education paper writing.

3.2 Designing Multi-Dimensional Academic Scoring Tasks

The academic evaluation of higher education papers is a complex cognitive process involving multiple different but interrelated dimensions. In order to achieve a more comprehensive evaluation of the quality of academic writing, this paper refines the scoring task into four core dimensions: logic, citation norms, argument rigor, and language quality. This multi-dimensional scoring design is not only closer to the evaluation criteria of teachers' scoring in actual teaching but also helps to improve the comprehensive ability of the model in semantic understanding, structure recognition, and language normative judgment.

At the model implementation level, this paper models each scoring dimension as an independent learning task and uniformly constructs it through a multi-task learning framework. The specific approach is to set up a dedicated output layer for each scoring dimension based on the shared pre-trained language model BERT. BERT, as the underlying encoder, is responsible for generating semantic representations, while different scoring dimensions are predicted through task-specific output heads. This design enables the model to share an overall understanding of text semantics while performing differentiated feature extraction and scoring strategy optimization for different tasks. By simultaneously modeling in multiple dimensions, it can not only improve the accuracy of each scoring task but also achieve information complementarity and collaborative optimization between tasks during the scoring process.

In terms of data, this paper constructs a scoring dataset covering multiple disciplines, including education, computer science, economics and other professional directions. All paper samples are manually scored and annotated in multiple dimensions, as shown in Table 1:

Table 1. Dataset subject distribution table

Discipline	Sample Size	Avg. Word Count	Annotation Ratio (Logic/Citation/Argument/Language)
Education	1200	3850	28%/22%/25%/25%
Computer Science	980	4200	30%/20%/30%/20%
Economics	850	3950	25%/25%/25%/25%

To ensure the quality of annotation, the entire annotation process follows a unified scoring standard and is reviewed by experts with teaching experience to minimize subjective bias. The multidisciplinary data setting also provides a basis for the generalization ability of the model, enabling it to maintain stable performance when facing texts with different writing styles and academic norms.

3.3 Implementing Knowledge Sharing in a Multi-Task Learning Framework

Multi-task learning allows tasks to share some knowledge of the model by jointly optimizing the objectives of multiple tasks. Specifically, this paper builds a multi-task

learning framework based on the BERT model, in which the underlying feature representations of all scoring tasks share the pre-trained semantic representation from BERT, while each task has its own independent output layer. Assuming that the output layer of each task is $\hat{y}_i$, the final multi-task loss function formula is (2):

$$\mathscr{L}_{MTL} = \sum_{i=1}^{M} \mathscr{L}_i(\hat{y}_i, y_i) + Reg(\theta) \quad (2)$$

Among them, $\mathscr{L}_i(\hat{y}_i, y_i)$ represents the loss function of the i-th task; $\hat{y}_i$ is the predicted value of the task; y_i is the true label; $Reg(\theta)$ is the regularization term; and θ is the parameter of the model. By sharing the underlying semantic information, the model can borrow knowledge from different tasks in the process of learning each scoring dimension, thereby improving the performance of the overall scoring system.

In the design of this paper, the BERT model is used as a shared underlying encoder to transform the input paper text into a set of high-dimensional semantic vectors through its deep context representation ability. The scoring dimensions of each task (such as logic, citation norms, argument rigor, and language quality) rely on this shared semantic representation and are predicted separately through independent output layers. In this process, BERT's ability to understand text is maximized, while the output layer of each scoring dimension focuses on the scoring criteria unique to that dimension. In this way, the model is able to handle multiple scoring tasks simultaneously, thereby improving the prediction accuracy on each task.

In addition, the multi-task learning framework can effectively avoid the overfitting problem. Since the model needs to optimize on multiple tasks at the same time, it must learn a more general feature representation instead of relying solely on the specific pattern of a certain task. This training strategy helps improve the generalization ability of the model on unseen data, enabling it to maintain high scoring accuracy and consistency in papers of different disciplines or writing styles.

3.4 Introducing Data Enhancement to Optimize Model Generalization Capabilities

Data augmentation is the process of processing the original training data to generate diverse samples so that the model can learn under more diverse input conditions. Text data usually has the characteristics of a single structure and fixed language patterns, which makes the trained model prone to overfitting on specific types of data. The goal of the training process is to minimize the following (3) loss function:

$$L_{total} = \frac{1}{N'} \sum_{i=1}^{N'} \mathscr{L}(f(x_i'), y_i) \quad (3)$$

Among them, $\mathscr{L}$ represents the loss function of the model, and $f(x_i')$ represents the score predicted by the model for the enhanced sample x_i'. Especially in academic writing, papers from different disciplines often have large differences in writing style and expression. Therefore, relying solely on the original training data may not cover all possible writing variants. In this case, data augmentation methods help the model better adapt to complex language structures and diverse writing styles by simulating various writing variants.

In this paper, several common text data enhancement techniques are adopted, including synonym replacement, sentence rearrangement, text perturbation, and random deletion. In terms of synonym replacement, this paper finds words that can be replaced in the sentence and randomly replaces them with their synonyms, thereby generating sentences with the same semantics but different expressions..

4 Results and Discussion

4.1 Experimental Analysis

(1) Multi-dimensional scoring accuracy evaluation experiment

This experiment aims to evaluate the performance differences of four scoring models in different academic dimensions of higher education paper writing. The four dimensions of logical structure, citation norms, argument strength and language quality are selected as scoring criteria, and the accuracy of the four methods of TF-IDF+SVM, LSTM, BERT single-task model and BERT combined with multi-task learning are compared. The experiment uses a uniformly annotated paper data set to ensure that different models are trained and tested under the same conditions. To improve the credibility of the results, each model is evaluated independently in each dimension, and Accuracy is used as the main evaluation indicator. The difference in scoring ability of each model in different dimensions is shown in a line chart. The specific data is shown in Figure 2:

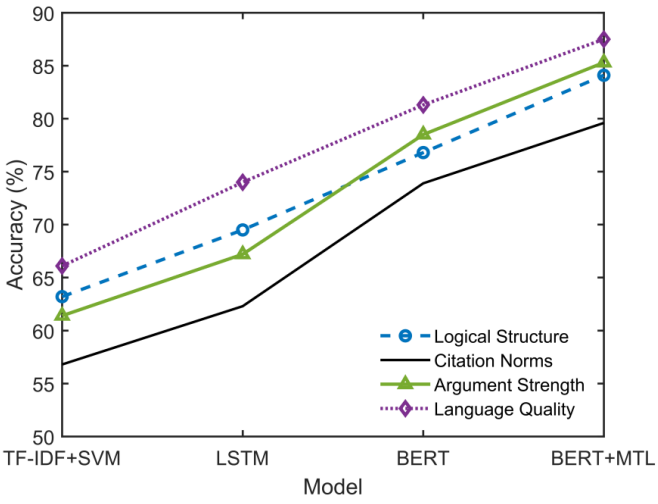

Fig. 2. Multi-dimensional scoring accuracy evaluation

In the multi-dimensional scoring accuracy evaluation experiment, the four models perform significantly differently in the four dimensions of "logic structure", "citation norms", "argument strength" and "language quality". The accuracy of TF-IDF+SVM in

each dimension is 63.2%, 56.8%, 61.4% and 66.1%, respectively, which is relatively low overall. The LSTM model is relatively good in "language quality" (74.0%), but has limited performance in the "citation norms" dimension (62.3%). The overall accuracy of the BERT single-task model is between 73.9% and 81.3%, which is a significant improvement. The BERT+ multi-task learning model proposed in this paper has reached 84.1%, 79.6%, 85.3% and 87.5% in four dimensions, respectively. The experimental results show that multi-task learning has a significant effect in improving the scoring system's ability to capture complex academic features, and the scoring accuracy and consistency are better than other models.

(2) Scoring error analysis experiment

The scoring error analysis experiment aims to evaluate the ability of different models to fit real human scores in academic scoring tasks. The experiment selects four core dimensions: logic structure, citation norms, argument strength, language quality, and calculates the MSE of each model in each dimension to quantify the difference between the predicted results and the actual scores. The experiment compares four models, including the traditional TF-IDF+SVM, LSTM, BERT single-task model, and BERT combined with multi-task learning model. By drawing a bar chart, the error performance of each model in different dimensions is intuitively displayed. Specifically, it is shown in Figure 3:

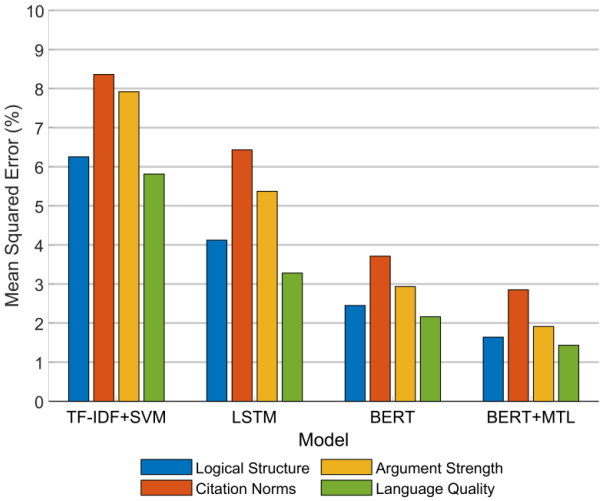

Fig. 3. Scoring error analysis evaluation

The results of the scoring error analysis experiment show that there are significant differences in the fitting ability of the four models in each academic dimension. The TF-IDF+SVM model has the highest mean square error in the "citation norms" dimension, reaching 8.36%, and the errors in the other dimensions are also above 5%. The LSTM model performs relatively well in the "language quality" dimension, with an error of 3.28%. The BERT single-task model performs stably overall, with MSEs of all

four dimensions below 4%, including 2.45% in the "logical structure" dimension. In contrast, BERT combined with a multi-task learning model achieves the lowest error in all dimensions, including 1.43% for "language quality" and 1.91% for "argument strength". This result shows that multi-task modeling has obvious advantages in improving the accuracy of scoring fitting, especially when dealing with complex structured language and deep semantics.

(3) Rating consistency comparison experiment

The rating consistency comparison experiment aims to evaluate the consistency between different models and manual scoring results in various academic dimensions. The experiment uses Cohen's Kappa coefficient as the consistency evaluation indicator, sets four dimensions: logical structure, citation norms, argument strength, and language quality, and compares the matching degree between the scoring results of the four models and manual scoring in classification, so as to reflect the scoring stability and reliability of each model in practical application. The specific details are shown in Figure 4:

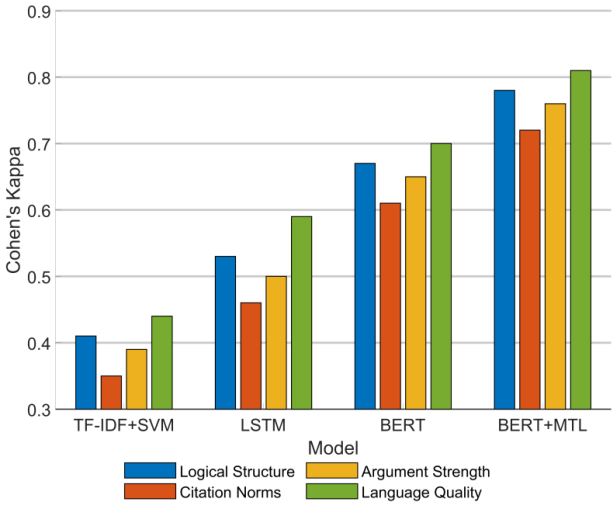

Fig. 4. Comparative evaluation of scoring consistency

The results of the scoring consistency comparison experiment show that there are obvious differences in the consistency between different models and manual scoring in each dimension. The TF-IDF+SVM model has the lowest consistency, and the Cohen's Kappa coefficient does not exceed 0.45 in any of the four dimensions, with the lowest value of 0.35 in the "citation norm" dimension. In the dimension of "language quality", the LSTM model reaches 0.59, slightly higher than other dimensions. The BERT single-task model achieves 0.67 and 0.65 in the dimensions of "logical structure" and "argument strength", respectively, showing good semantic understanding ability. BERT combined with the multi-task learning model performs the most stably, reaching the highest value of 0.81 in the "language quality" dimension and also achieves a

relative advantage of 0.72 in the "citation norm" dimension. Overall, the consistency of the model in all dimensions is better than other comparison methods, indicating that its scoring is closer to manual judgment and has stronger stability and credibility.

(4) Generalization ability test experiment

The generalization ability test experiment introduces a subset of papers from three different disciplines (education, computer science, and economics) to evaluate the performance of four scoring models in non-training set fields. The experiment uniformly uses accuracy as a measurement standard to examine the stability and adaptability of each model in cross-domain scenarios, thereby reflecting the strength of its generalization ability in real applications. The specific data are shown in Table 2:

Table 2. Generalization ability test evaluation

Model	Education Subset	Computer Science Subset	Economics Subset	Average Accuracy
TF-IDF + SVM	61.4	58.9	60.7	60.3
LSTM (Single-task)	68.2	65.3	66.8	66.8
BERT (Single-task)	77.1	73.8	75.6	75.5
BERT + Multi-task	83.5	81.2	82.7	82.5

The experimental results show that there are significant differences in the scoring performance of different models on each subset. The accuracy of the TF-IDF+SVM model on the three subsets is less than 62%, with an average of only 60.3%, and the generalization ability is weak. The LSTM model performs well in the education subset (68.2%), but slightly declines in the computer direction (65.3%). The overall performance of the BERT single-task model is relatively stable, with the accuracy of the three subsets all above 73%, with an average of 75.5%. The accuracy of BERT combined with the multi-task learning model in various fields is 83.5%, 81.2% and 82.7%, respectively, with an average of 82.5%, which is significantly better than other models, indicating that this method has a strong ability to adapt to cross-disciplinary scoring.

5 Conclusion

This paper proposes a scoring method that combines the BERT model and multi-task learning to improve the accuracy, consistency and cross-domain adaptability of scoring, focusing on the academic automatic scoring task in higher education paper writing. By applying a multi-task modeling mechanism with four dimensions: logic, citation norms, argument rigor and language quality, the model can achieve collaborative scoring on multiple academic indicators, effectively breaking through the limitations of traditional methods on the deep semantic understanding of text. Experimental results show that the proposed method performs well in scoring accuracy, error control, and consistency with manual scoring, and has good generalization ability, making it suitable for automatic scoring of papers in different disciplines and university contexts. Although the model has achieved good results in multiple dimensions, this study still has certain limitations. Future research can be expanded in two directions: first, introducing teacher feedback and scoring explanation mechanisms to enhance the inter-

pretability of the model; second, combining more language model structures to explore automatic scoring systems for multi-language and multi-style scenarios, and further promote their implementation in actual educational evaluation.

References

1. Messer M, Brown N C C, Kölling M, et al. Automated grading and feedback tools for programming education: A systematic review[J]. ACM Transactions on Computing Education, 2024, 24(1): 1-43.
2. Ramesh D, Sanampudi S K. An automated essay scoring systems: a systematic literature review[J]. Artificial Intelligence Review, 2022, 55(3): 2495-2527.
3. Hahn M G, Navarro S M B, Valentín L D L F, et al. A systematic review of the effects of automatic scoring and automatic feedback in educational settings[J]. Ieee Access, 2021, 9(1): 108190-108198.
4. Kooli C, Yusuf N. Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading[J]. International Journal of Human–Computer Interaction, 2025, 41(5): 3388-3399.
5. Combéfis S. Automated code assessment for education: review, classification and perspectives on techniques and tools[J]. Software, 2022, 1(1): 3-30.
6. Wilson J, Huang Y, Palermo C, et al. Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write[J]. International Journal of Artificial Intelligence in Education, 2021, 31(2): 234-276.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.