# A Review of the Application and Progress of Deep Learning in Automatic Essay Grading Systems

**Yitian Zhang**

School of Computer Science, Xinjiang Normal University, Urumqi, Xinjiang Uyghur Autonomous Region 830054

**Abstract:** ***With the rapid development of technology, deep learning technology has demonstrated excellent application effects in many fields due to its powerful feature learning and representation capabilities. In the automatic essay grading system, the application of deep learning is becoming increasingly important. This article aims to comprehensively review the latest applications and significant progress of deep learning in automatic essay grading in recent years. The current automatic essay grading technology still faces some challenges, such as the accuracy of cross language and cross-cultural grading, as well as the stability and reliability of the grading system. We have conducted a thorough analysis of these issues and proposed possible solutions. Looking ahead to the future, we expect deep learning technology to play a greater role in the field of automatic essay grading, promoting the intelligence and precision of educational evaluation.***

**Keywords:** Deep learning; Automatic grading of essays; Natural language processing; Artificial intelligence.

## 1. INTRODUCTION

Composition is the core of language education, reflecting students' language abilities, and its grading methods have evolved with technological progress. Traditional grading relies on teachers' subjective judgment, which is meticulous but has inconsistent standards and low efficiency. Therefore, researchers explored computer automated scoring. Deep learning technology brings new opportunities for essay grading, as it can automatically learn and extract deep features, providing more objective and accurate scoring criteria.

This article reviews the current status of manual feature and deep learning essay grading techniques, analyzes the advantages and disadvantages of each method and their practical application effects, and focuses on challenges such as grading fairness, accuracy, and adaptability to cross linguistic and cultural backgrounds. We hope to provide reference for future researchers, promote technological improvement, assist in the intelligence of educational evaluation, and support personalized learning for students. We hope to achieve more scientific and efficient essay evaluation through continuous technological advancements, ultimately enhancing students' writing abilities. Xiangyu et al. [1] developed a novel granule extrusion-based 3D printing method for polyolefin elastomers (POE), employing response surface methodology to optimize mechanical properties. This technological innovation aligns with broader digital transformation trends examined by Chen et al. [2], who quantitatively analyzed the green innovation effects of the digital economy using advanced econometric models. Logistics optimization has seen substantial progress through AI implementation. Meng et al. [3] proposed a deep learning framework for green warehousing logistics, simultaneously addressing site selection and path planning challenges. These operational improvements complement Wang's [7] Bayesian optimization approach for urban delivery network reconfiguration, demonstrating how AI can enhance supply chain efficiency at multiple levels. In healthcare, Wang et al. [4] conducted groundbreaking work mapping the immune microenvironment in gastrointestinal cancers, providing a cellular atlas that could inform future immunotherapy development. Complementing this diagnostic advancement, Li [8] developed machine learning systems for enhanced adverse event monitoring in Phase IV clinical trials, showcasing AI's growing role in pharmaceutical safety. Urban informatics has benefited from several AI-driven innovations. Li et al. [5] pioneered gamification techniques for smart city data visualization, while their subsequent work on named entity recognition [9] improved real-time processing of urban data streams. These developments in civic technology are being paralleled by commercial applications, as evidenced by Song's [6] intelligent demand forecasting systems and their later research on AI-enhanced internal tools [11] for e-commerce operations. Financial technology has also advanced through AI applications, with Yang [10] demonstrating the effectiveness of LightGBM algorithms in analyzing China's complex stock market dynamics.

## 2. ESSAY GRADING METHOD BASED ON MANUAL FEATURES

The manual feature-based essay grading method was the main research direction in the early days of automated essay grading. The core of this method is to manually design and extract a series of features that can reflect the quality of the composition based on the grading standards and requirements. These features may include but are not limited to richness of vocabulary, correctness of grammar, rationality of structure, and depth and breadth of content.

In early research, L. M. Rudner and T. Liang et al. (2002) [1] attempted to use Bayesian classifiers and K-nearest neighbor methods to classify essays. They first classify the essays into several categories based on their different characteristics, then further integrate the complexity features of the text based on the classification results, and finally use regression methods to score the essays. Although this method was simple, it opened up new ideas for automatic essay grading at that time.

Benjamin B. Bederson et al. (2003) [2] also followed a similar approach, using Bayesian classifiers to achieve essay grading. Their research further validated the feasibility of a manual feature-based essay grading method. However, this method also has certain limitations, as the selection and design of features largely rely on the researcher's experience and intuition, which may not fully and accurately reflect the true level of the composition.

To overcome this limitation, Dan Bikel et al. (2010) [3] proposed a method of scoring essays by weighting and averaging multiple eigenvalues. They comprehensively considered multiple aspects of the composition, including vocabulary, grammar, structure, etc., assigned corresponding weights to each feature, and calculated the weighted average to obtain the final score. This method has improved the accuracy and comprehensiveness of scoring to a certain extent.

As research progresses, researchers begin to explore more refined and complex scoring methods. Keisuke Sakaguchi et al. (2015) [4] introduced a pair wise ranking method, which utilizes support vector machines (SVM) for essay grading. This method not only considers the relative advantages and disadvantages between compositions, but also automatically learns and adjusts the grading model through machine learning algorithms, thereby further improving the accuracy of grading.

Chen et al. (2013) [5] went further by using a list wise sorting method. They trained a ranking model using the LambdaMART algorithm, which can simultaneously consider the relative order relationship between multiple essays, thus more accurately evaluating the quality of essays. This method achieved significant results at that time and provided important references for subsequent research on automatic essay grading.

In addition, Zhao Wei et al. (2020) [6] also attempted to use domain adaptation techniques to solve the problem of cross topic essay grading. They realized that there may be significant differences in content and style among essays on different topics, so domain adaptation techniques are needed to adjust the scoring model to meet the needs of different topics. This method provides a wider applicability for the practical application of automatic essay grading technology.

Overall, the manual feature-based essay grading method laid a solid foundation for the development of automatic essay grading in the early days. With the rise and development of deep learning technology, the field of automatic essay grading has also ushered in new opportunities and challenges. Deep learning technology can automatically learn and extract deep features from essays, providing more objective and accurate basis for essay grading, and is expected to become the mainstream method in the field of automatic essay grading in the future.

## 3. ESSAY GRADING METHOD BASED ON DEEP LEARNING

In recent years, deep learning technology has demonstrated strong capabilities in multiple fields, including essay grading. With the continuous advancement of deep learning algorithms and models, more and more researchers are exploring their application in essay grading, in order to achieve more accurate and objective grading.

Hua Yu (2020) [7] conducted in-depth research on automatic essay scoring technology across prompt scenarios at Nanjing University. He constructed a deep learning model that can automatically grade essays based on their characteristics in different prompt scenarios. The highlight of this study is that its model has strong generalization ability and can adapt to essay grading needs in different scenarios. Through extensive experimental verification,

the model has achieved significant results in essay grading across prompt scenarios, providing a new solution for the flexibility and accuracy of essay grading.

Xiangqun Cheng (2022) [8]: At the University of Chinese Academy of Sciences, an automatic scoring system for high concurrency composition was designed. He addressed the performance bottleneck that traditional essay scoring systems may encounter when dealing with a large number of essays, and improved the performance and stability of the scoring system by optimizing algorithms and system architecture. This study not only solves the problem of essay grading in high concurrency scenarios, but also provides strong technical support for the practical application of essay grading technology.

Yong Yang et al. (2021) [9]: Proposed multiple technological innovations in the field of automatic essay grading at Xinjiang Normal University. They first developed the MLSF model, which effectively integrated multi-level semantic features by combining CNN and hybrid neural networks, significantly improving the scoring performance on the Kaggle ASAP dataset, with QWK reaching 79.17%. Subsequently, in response to the issues of long text processing and topic relevance, they introduced the TASE model, which utilizes multi head attention and BERT models to deeply analyze semantics and further optimize scoring performance. The relevant code and dataset have been open sourced, providing valuable resources for the academic and educational communities.

In addition to the aforementioned researchers, Ping Xiao (2007) [11], Wenjuan Li (2023) [12], Yiyi Liao (2022) [13], Lin Zhang (2022) [14], Yifan Wang (2022) [15], Yinxin Yang (2020) [16], Meizhen Wang (2021) [17], and others have also conducted in-depth research on essay grading techniques from different perspectives. They either start with the language features of the composition, focus on the structure and content quality of the composition, or explore the best application of deep learning models in composition grading, providing multi-dimensional and comprehensive support for the development of composition grading technology.

Through continuous research and improvement, we have reason to believe that this technology will play an increasingly important role in future educational assessments, providing stronger support for personalized learning and development of students.

## 4. CURRENT PROBLEMS AND CHALLENGES

Although deep learning has achieved significant results in automatic essay grading, it still faces some challenges and problems in practical application. These issues mainly focus on data sparsity, interpretability of models, and uniformity and objectivity of scoring criteria.

Data sparsity is an important issue. The training of deep learning models relies on a large amount of annotated data, however, in the field of essay grading, obtaining annotated data is not easy. On the one hand, grading essays requires a significant amount of manpower and time costs, and the quality of annotation is limited by the professional level and subjective judgment of the annotator. On the other hand, due to the diversity and complexity of essays, even the same essay may receive different scores under different grading criteria, which further increases the difficulty of data annotation. Data sparsity not only affects the training effectiveness of the model, but may also lead to overfitting and reduce the accuracy of the scoring. The interpretability of models is also a current challenge. Deep learning models typically have complex network structures and a large number of parameters, making the learning and decision-making processes of the model difficult to explain. In essay grading, users often want to understand the specific basis and reasons for the grading results, in order to better understand the strengths and weaknesses of the essay and make improvements. However, current deep learning models are difficult to provide such explanatory information, which to some extent limits their widespread application in essay grading. The standardization and objectivity of scoring criteria is also an urgent issue that needs to be addressed. Essay grading involves multiple aspects of evaluation, including language expression, content quality, and discourse structure. Different scoring criteria may lead to inconsistent scoring results and even controversy. In addition, due to the subjectivity of essay grading, different raters may give different scores to the same essay. Therefore, how to establish a unified and objective scoring standard to reduce the influence of subjective factors is an important issue that needs to be addressed in current automatic essay grading technology.

To address the aforementioned issues, researchers are constantly exploring and innovating. For example, to address the issue of data sparsity, unsupervised learning or semi supervised learning methods can be attempted to utilize unlabeled data for pre training and improve the model's generalization ability. At the same time, it is also possible to consider introducing transfer learning strategies to transfer knowledge learned on one dataset to other

related datasets, thereby accelerating model training and improving performance. In terms of model interpretability, research can be conducted on how to design more transparent deep learning models or develop auxiliary tools to provide explanatory information for rating results. In addition, the development of unified and objective scoring standards also requires the joint efforts of experts in the fields of education and computer science to continuously improve and optimize the scoring standards through in-depth research and practice.

## 5. FUTURE DEVELOPMENT DIRECTION AND OUTLOOK

With the continuous advancement of deep learning technology and changing demands in the field of education, automatic essay grading technology will also usher in new development opportunities. In the future, the development direction of deep learning in automatic essay grading includes cross linguistic and cross-cultural essay grading, personalized and adaptive grading systems, and in-depth exploration of the integration of technology and education. Cross linguistic and cross-cultural essay grading is currently an important development direction. With the acceleration of globalization and the trend of internationalization of education, more and more students need to undergo educational assessments in different languages and cultures. Therefore, developing deep learning models that can score essays across languages and cultures has important practical significance. This will help students better adapt to an international educational environment and enhance their cross-cultural communication skills. Personalized and adaptive scoring systems are also an important development direction for future automatic essay scoring technology. Each student's writing level and style are different, so the grading system should be able to provide personalized evaluation and feedback based on the student's actual situation. By introducing technologies such as user profiling and learning path analysis, a more personalized and adaptive scoring system can be constructed to provide students with more accurate and effective writing guidance.

## 6. CONCLUSION

After comprehensively reviewing the application and latest progress of deep learning in automatic essay grading systems, it is not difficult to find the enormous potential and broad prospects that deep learning has shown in this field. With the continuous advancement of technology and the deepening of scientific research, we have ample reason to expect deep learning to play a more central role in automatic essay grading in the future. Its powerful data processing and analysis capabilities make essay grading more accurate and objective, providing an extremely intelligent and efficient auxiliary tool for writing teaching. However, the development of technology always comes with challenges.

## REFERENCES

[1] Xiangyu, G., Yao, T., Gao, F., Chen, Y., Jian, X., & Ma, H. (2024). A new granule extrusion-based for 3D printing of POE: studying the effect of printing parameters on mechanical properties with “response surface methodology”. Iranian Polymer Journal, 1-12.
[2] Chen, K., Zhao, S., Jiang, G., He, Y., & Li, H. (2025). The Green Innovation Effect of the Digital Economy. International Review of Economics & Finance, 103970.
[3] Meng, Q., Wang, J., He, J., & Zhao, S. (2025). Research on Green Warehousing Logistics Site Selection Optimization and Path Planning based on Deep Learning.
[4] Wang, Y., Yang, T., Liang, H., & Deng, M. (2022). Cell atlas of the immune microenvironment in gastrointestinal cancers: Dendritic cells and beyond. Frontiers in Immunology, 13, 1007823.
[5] Li, X., Wang, J., & Zhang, L. (2025). Gamifying Data Visualization in Smart Cities: Fostering Citizen Engagement in Urban Monitoring. Authorea Preprints.
[6] Song, X. (2025). Improving User Experience in E-commerce Through Intelligent Demand Forecasting and Inventory Visualization.
[7] Wang, J. (2025). Bayesian Optimization for Adaptive Network Reconfiguration in Urban Delivery Systems.
[8] Li, T. (2025). Enhancing Adverse Event Monitoring and Management in Phase IV Chronic Disease Drug Trials: Applications of Machine Learning.
[9] Li, X., Wang, J., & Zhang, L. (2025). Named entity recognition for smart city data streams: Enhancing visualization and interaction. Authorea Preprints.
[10] Yang, J. (2025). Application of LightGBM in the Chinese Stock Market.
[11] Song, X. (2025). User-Centric Internal Tools in E-commerce: Enhancing Operational Efficiency Through AI Integration.