



---

# IMPROVING THE FAIRNESS AND DEPTH OF AUTOMATED ESSAY SCORING SYSTEMS THROUGH ADVANCED MACHINE LEARNING TECHNIQUES

---

Fatimah Nowh Aljaroudi - 242000474

MIDOCEAN UNIVERSITY – MASTER OF AI

## Contents

|                           |    |
|---------------------------|----|
| Introduction .....        | 2  |
| Problem Statement.....    | 2  |
| Research Objectives:..... | 3  |
| Literature Review .....   | 3  |
| Conclusion.....           | 9  |
| Bibliography: .....       | 10 |

## Introduction

Automated Essay Scoring (AES) systems have improved over the years, from using simple manual tools in the 1960s to advanced systems that used natural language processing (NLP) and machine learning (ML). The developments started by employing simple tasks like word counting, and grammar analyzing, currently integrated AES technologies evaluate essays on different areas, such as context organization and essay theme. These new technologies helped to improve the efficiency in AES effectively. Despite the significant advancements, AES techniques are struggling at effectively understanding human expression writing. Thus, ASE systems are facing many challenges like the lack of training data, resulting potential systematic

## Problem Statement

During the past few years, Automated Essay Scoring (AES) has significantly enhanced from basic evaluation such as word counting to sophisticated metrics using Natural Language Processing (NLP) and Machine Learning (ML). However, current AES systems still face substantial challenges, especially in evaluating the full meaning of human expressions, like evaluating argument strength, creativity, and nuanced content. These limitations are often encountered because of the inadequate training datasets, which leads to potential systematic biases and over-reliance on certain features. These issues compromise the fairness and accuracy of AES, affecting academic grading and student feedback. Current challenges highlight the need for further research to develop more advanced AES systems that can handle different academic contexts and content forms effectively.

## Research Objectives:

- To analyze the effect of the diversity of training dataset on the performance and fairness of AES systems.
- To evaluate the effectiveness of existing AES technologies in processing different forms of assignments and feedback, focusing on strengths and weaknesses.
- To develop a new ML model that includes a diverse range of linguistic and thematic elements to enhance the understanding and scoring complex assignments.
- To propose improvements in current AES models that reduce biases and over-reliance on non-essential features.
- To validate the proposed ML model against traditional and current AES systems through thorough testing with diverse assignments datasets.

## Literature Review

| Publication Year | Author(s)                  | Methodology                                       | Key Findings                                                                              | Weaknesses/Limitations                            |
|------------------|----------------------------|---------------------------------------------------|-------------------------------------------------------------------------------------------|---------------------------------------------------|
| 2023             | D. Ifenthaler et al.       | Review of AES technologies, NLP, Machine Learning | Evolution of AES, advancements in NLP and machine learning, identification of limitations | Challenges in depth of analysis, potential biases |
| 2020             | Kumar, V.S., Boulanger, D. | Deep learning models on ASAP dataset              | High QWK scores, improved                                                                 | Limited dataset diversity                         |

|      |                                       |                                                       |                                                                                                                                                                                                 |                                                                                                |
|------|---------------------------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|
|      |                                       |                                                       | interpretability of AES                                                                                                                                                                         |                                                                                                |
| 2022 | Weegar, Idestam-Almquist              | Two-step ML grading with manual review and clustering | 64-74% reduction in grading workload                                                                                                                                                            | High dependency on manual review for accuracy                                                  |
| 2024 | Xiao et al.                           | Experiments with LLMs                                 | Effective use of LLMs in scoring essays                                                                                                                                                         | Similar content, limited new insights                                                          |
| 2021 | Masaki Uto                            | Comprehensive review of DNN-based AES models          | DNN-AES models offer an effective alternative to traditional AES by eliminating the need for manual feature engineering. The review categorizes and details models for four types of AES tasks. | The paper does not specify weaknesses or limitations.                                          |
| 2022 | Sebastian Gombert, Aron Fink, Tornike | Controlled study comparing AI-driven                  | AI-driven feedback was more helpful and enhanced                                                                                                                                                | Challenges in accurately interpreting diverse responses; variability in feedback reception and |

|  |                                                                                                     |                                                                                                                    |                                                                                       |                                                  |
|--|-----------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|--------------------------------------------------|
|  | Giorgashvili, Ioana Jivet,<br>Daniele Di Mitri,<br>Jane Yau,<br>Andreas<br>Frey, Hendrik<br>Drachse | feedback<br>(using NLP<br>models GBERT<br>and T5) with<br>traditional<br>feedback<br>methods on<br>student essays. | student<br>reflection and<br>understanding<br>compared to<br>traditional<br>feedback. | dependence on<br>comprehensive training<br>data. |
|--|-----------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|--------------------------------------------------|

The book chapter written by D. Ifenthaler elaborated this particular area. The chapter aimed to go through; the evolution, operational mechanisms, and current challenges of automated essay scoring systems. They were designed to automate the evaluation of text-based essays. To achieve this goal, the writer reviewed AES technologies and methodologies, focusing mainly on improvements in natural language processing (NLP) and machine learning (ML) which develop the capability of such systems to evaluate and grade the written content.

The chapter addresses three important findings. First, AES systems have enhanced significantly since the 1960s, starting from first systems such as Project Essay Grader (PEG) that focused on basic metrics like counting words and grammar check. Also, current developments have engaged more complicated NLP and ML technologies, allowing for a more detailed analysis of assignments. For example: content, style, organization, and techniques. In spite of these advancements, AES systems still facing issues related to the ability of capturing the full depth and nuance of human viewpoints, especially in argument strength, creativity, and deeper content analysis.

The author effectively outlines the advantages of AES systems, which includes their capability to reduce the work load of grading and provide effective feedback. On the other hand, it critically highlights the weaknesses as well, including the disadvantage of over-reliance on non-

important features and some potential of systemic biases. This would be encountered if the training data are not sufficient enough. (Zawacki-Richter & Jung, n.d.)

The research paper written by Vivekanandan S. Kumar & David Boulanger presents a detailed exploration of Automated Essay Scoring (AES), by operating deep learning technologies. The main goal of this paper is to investigate the accuracy and effectiveness of automated essay scoring according to a

grading rubric. The method used was introduces dataset of a manual graded essay, then applying seep neural networks to predict rubric scores. The deep learning model showed a higher QWK than the previous AES systems, including a significant enhancement comparing with human assessment.

The study focused on the interpretability of AES systems, proposing that transparency in how grades are measured could improve the acceptability and reliability of automated scoring, which elaborate the novelty of this study.

The strength of this scientific paper is that it provides a detailed analysis automated essay scoring by using significant dataset. However, the limitations of using a single source for the dataset considered as a weakness for this paper.(Kumar & Boulanger, 2021)

‘Reducing Manual Grading Workload through Machine Learning Techniques’ article investigates the major load of manual grading in educational field, particularly in computer science courses.

The main goal of this study is to determine the integration of machine learning (ML) to reduce the workload of manual regular grading, without reducing the quality of grading.

The method used by the researchers involves a two-step automated grading process, the initial step of the automated grading process is collecting a subset of a manual graded answers as a data to train ML models. Then the trained ML models start to grade the remaining of student answers. This article achieved a workload reduction by 64% to 74% compared to manual grading, increasing the efficiency and accuracy of grading.

Examining the practical application of ML in educational grading, providing detailed analysis of different strategies to implement automated grading system, are considered as a strength of the essay. While the limitation of this study that it mostly relies on the accuracy of ML models, and it can't be generalized on all course contexts.(Weegar & Idestam-Almquist, 2024)

Xiao and his colleagues examine specifically how the two models of GPT-3.5 and GPT-4 can improve the efficiency and accuracy of grading by comparing novice and expert graders. The objective of this article is the evaluation of the effectiveness of both GPT-3.5 and GPT-4 in automating scoring process using the experiment method in this essay to examine LLMs on datasets and expert graders. This essay emphasizes the significant enhancement of grading accuracy and efficiency, illustrating two potentials solutions, standalone grading and support grading system to support humans.

Moreover, it is possible to summarize the strengths and weaknesses into two main points to each one. The strengths are, detailed empirical analysis and incorporation of human-AI interpretations, while the weaknesses are, limited demographic focus and potential bias in AI interpretations.(Changrong Xiao et al., 2024)

"A Review of Deep-Neural Automated Essay Scoring Models" article reviews various deep learning models used in automated essay scoring, mainly concentrating on their effectiveness across different grading tasks. The main objective of this article is to review different deep learning models used in automatic essay scoring, using critical reviewing the literature and absorbing on the effectiveness of Deep Neural Networks (DNNs) and Recurrent Neural Network (RNNs). It is also illustrated the promising future in the efficiency and accuracy of essay scoring using the DNNS.

The Comprehensive overview and critical analysis of DNNs and RNNs models specifically, are the strengths for this article. On the other hand, the weaknesses are, potential reliability issues and confined to a limited educational format. (Gombert et al., 2024)

'Systematic Review of AI-Driven Feedback in Educational Psychology' cases study implementing an AI-driven feedback system for German-speaking university students in an educational

psychology course. The system technology used is the Natural Language Processing (NLP), GBERT and T5 models. It is used to provide detailed and personalized feedback on student assignments regarding in related to the learning tips provided to the students. The effectiveness of this evolutionary AI-driven feedback was compared with assessor's feedback methods. It has been shown that the new technology was recognized as more helpful by students, enhancing their understanding and reflective practices. (Uto, 2021)

The study illustrates the ability of AI to deliver personalized and effective feedback. However, and similar to other studies, challenges are the AI's protentional to interpret different student responses accurately and the variability in feedback reception.

The main important challenge addressed on AES systems reviews is dealing with scoring rubric ambiguity and grading subjectivity. This study suggests that the advances NL processing capabilities of the LLMs would solve this issue via providing trusted and personalized feedback.

The exploration of Automated Essay Scoring (AES) systems across the studies reviewed in this essay shows the increase demands on the advanced technologies including ML, deep learning, and Large Language Models (LLMs) to develop the efficiency of essay grading. They provided a detailed understanding of how these methods enhancements have reform the grading systems. This was mostly elaborated on both D. Ifenthaler's historical and operational analysis of AES and the application of deep learning models in AES as detailed by Vivekanandan S. Kumar & David Boulanger.

Although the studies demonstrated the importance of AES to educational assessment revolutionary, they also discussed multiple challenges. This can be summed up on one main issue which are the biases encountered due to inadequate training datasets, over-reliance on specific features, and the generalizing across multiple educational criteria. Furthermore, future improvements would be needed on areas like ML accuracy and training data diversity.

## Conclusion

In conclusion, the benefits of integrating AES into educational systems would be summarized on reducing instructors' workload and providing timely, personalized feedback. However, it is suggested the need of balancing such technologies with human oversight to guarantee the fairness and comprehensiveness in grading.

## Bibliography:

- Changrong Xiao, Wenxing Ma, Sean Xin Xu, & Kunpeng Zhang. (2024). *From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape*. Beijing Xicheng Educational Research Institute.
- Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the Automated Assessment of Student Essay Content to Highly Informative Feedback: a Case Study. *International Journal of Artificial Intelligence in Education*.  
<https://doi.org/10.1007/s40593-023-00387-6>
- Kumar, V. S., & Boulanger, D. (2021). Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, 31(3), 538–584. <https://doi.org/10.1007/s40593-020-00211-5>
- Uto, M. (2021). A review of deep-neural automated essay scoring models. In *Behaviormetrika* (Vol. 48, Issue 2, pp. 459–484). Springer Japan. <https://doi.org/10.1007/s41237-021-00142-y>
- Weegar, R., & Idestam-Almquist, P. (2024). Reducing Workload in Short Answer Grading Using Machine Learning. *International Journal of Artificial Intelligence in Education*, 34(2), 247–273. <https://doi.org/10.1007/s40593-022-00322-1>
- Zawacki-Richter, O., & Jung, I. (n.d.). *Handbook of Open, Distance and Digital Education*.