INFORMATICS JOURNALS

66

SHRI-LEKH (Multi-disciplinary Research Journal) Vol 1(2), December 2025, p. 66-71

Online ISSN: 3139-5694
DOI:10.18311/SHRI-LEKH/2025/v1i2/53554

Unlocking the Potential of Machine Learning, Deep Learning and Nlp in Automated Grading of Essays: A Comprehensive Review

Sohini Banerjee and Shweta. M. Barhate*

Department of Electronics and Computer Science, RTM Nagpur University, Nagpur - 440033, Maharashtra, India

Abstract

Machine Learning mainly deals with the algorithms and models that can learn from training data and make predictions on unknown data without being explicitly programmed. Deep Learning is that subset of Machine Learning that deals with vast amount of data to create complex models. Natural Language Processing (NLP) works on language and speech data and understands how computers and human language interact while encompassing algorithms and models that allow computers to comprehend, decipher, understand, and also generate natural language. The concept of automatically grading essays, also known as Automated Essay Scoring (AES) or automated essay grading (AEG), involves assessing and evaluating essays or written responses by deploying various algorithms without human assistance. AEG has proved to be a significant milestone in the journey of evolution of artificial intelligence in educational arena, addressing the urgency for scalable, consistent, and timely assessment of written paragraphs, essays and compositions. This comprehensive study examines the evolution, methodologies, and recent scenario of automated essay grading systems, with special mention of ML, DL and NLP approaches. This research unveils the historical development from traditional rule-based systems through feature-engineering moving towards ML and DL models to contemporary transformer-based architectures and large language models emphasizing on the milestones of this process. This study systematically summarizes traditional ML techniques including support vector machines, random forests, and ensemble methods, alongside deep learning architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and attention-based transformers including BERT, GPT, and RoBERTa. This research review critically assesses all real-world applications and importance of Automated Essay Grading including its multi-faceted features. The results of this analysis suggests that while transformer-based models achieve outstanding performance, challenges remain in cross-prompt generalization, interpretability, and ethical deployment, necessitating hybrid approaches that combine the strengths of feature-based explainability with deep learning's modelling capacity.

Keywords: Assessment, Automated Essay Grading, BERT, Deep Learning, Educational Technology, Explainable AI, Machine Learning, Natural Language Processing, Transformers

INTRODUCTION

In the domain of education evaluation of huge volume of written essays is a resource-intensive task. While traditionally manual grading was encouraged but with the changing times and technological advancements along with massive enrolments and requirement of immediate feedback the necessity of automated essay grading system became inevitable.

Email: sohini.banerjee01021984@gmail.com
*shwetab73@yahoo.com

Automated Essay Grading

The practise of automatically grading essays, also known as Automated Essay Scoring (AES) or Automated Essay Grading (AEG) mainly involves grading,assessing and evaluating essays or written responses along with assignment of scores or feedback without the assistance of humanraters. Basically it extracts different features from the essay or composition and evaluates them based on assigned rubrics. Score prediction may vary from model to model like holistic scores, analytic or trait scores and often feedback and recommendations for improvement. Beginning with the classical models to the transformers and LLMs basically this is explored to speed up and automate the essay grading procedure, offering a more scalable, efficient and uniform method of evaluation. An essay's content, structure, coherence, language use, and grammar are all examined by automated grading systems employing statistical models, machine learning and deep learning algorithms. The methodology initiates with essay submission or a dataset of pre submitted essays, data preparation, feature extraction, application of a scoring model with the underlying algorithms, and finally output of grades or scores along with feedback. Most of the educational institutions and universities explore this process to speed up the grading procedure. To overcome the several disadvantages like capturing the intricacies of writing quality, validity of feature extraction and bias automated essay grading systems are still being improved.

Machine Learning and Deep Learning Algorithms

Since manual grading systems often suffers from high cost in terms of several working hours of expert labour, inter rater variability and long waiting hours for feedback due to slow evaluation different machine learning models have been deployed for AEG such as Decision Tree, Random Forest, SVM, Gradient Boosting, Linear, Ridge, Lasso and many more. Deep Learning facilitates black box architectures and helps in implementation of RNNs, CNNs, Hybrid models, Attention Layers and fine-tuned transformers. Since Deep Learning models learn richer representations by capturing long range dependencies and contextual meaning and remove the need of manual feature engineering, Deep Learning models are preferred over the Machine Learning ones.

Natural Language Processing

Due to the multi-dimensional capabilities of NLP such as tokenisation, POS tagging, syntactic parsing, grammatical error detection these systems can mimic human graders more efficiently to evaluate the linguistic nuances. Text parsing, named entity recognition, sentiment analysis, language translation, question answering, and text generation are all included in NLP activities. NLP also adds coherence modelling and thus helps language models, and deep learning architectures like recurrent neural networks and transformers to evolve with the help of NLP. AES is a method of evaluating educational performance hence different natural language processing (NLP) algorithms like Bag of Words, N grams, Latent Semantic Analysis and Word embedding are also used for evaluating scores and generating recommendations.

RESEARCH OBJECTIVE

Firstly this study shows the journey of Automated Essay Grading starting from the traditional feature engineering methods to the machine learning models followed by deep learning ones. Finally this journey reaches the current era of Large Language Models and Hybrid models to improve accuracy and precision of evaluation. The primary goal is to compare different models and analyse the pros and cons of each model so that the optimal model can be selected for deployment according to the requirement. The main emphasis is given to uncover the significance of these models so that both the students and the educators get benefitted by the exploration of these holistic approaches.

Significance of Automated Essay Grading using Machine Learningand NLP

Automated Essay Grading (AEG) utilizing machine learning and natural language processing (NLP) techniques has emerged as a significant area of

research with numerous implications and advantages. The adoption of AEG offers several key benefits and addresses various challenges associated with traditional manual grading methods. The following are the significant aspects of AEG using machine learning and NLP, (Figure 1).

With the massive enrolments in educational institutions AEG significantly reduces the evaluation time and effort required for grading large volumes of essays manually and thus providing educators with more time for other critical tasksby automating the grading process.Moreover due to personal human biases and difference in backgrounds manual grading often shows high variations whereas AEG offers consistency and objectivity in the evaluation process as Machine Learning and Deep Learning algorithms follow a standardised approach to the complete dataset. This ensures reduction of error leading to uniform grading across all essays. ML evaluates all essays based on the same criteria and hence ensures consistency and objectivity in the grading process of AEG. Automation of this system also provides immediate and prompt feedback by predicting scores and comments for improvement by evaluating their strengths and weaknesses. Online learning environments make the application more scalable by breaking the physical boundaries between the students and the evaluators. These systems make the whole process accessible to large number of people spread across the world enabling them receive immediate score and feedback. Deployment of NLP techniques can reveal valuable information about students writing skills, language proficiency, and conceptual understanding and thus help in systematic analysis of the essays. AEG can further support and enhance pedagogical practices by providing valuable insights to educators. Age old technique of manual grading requires significant resources, including man

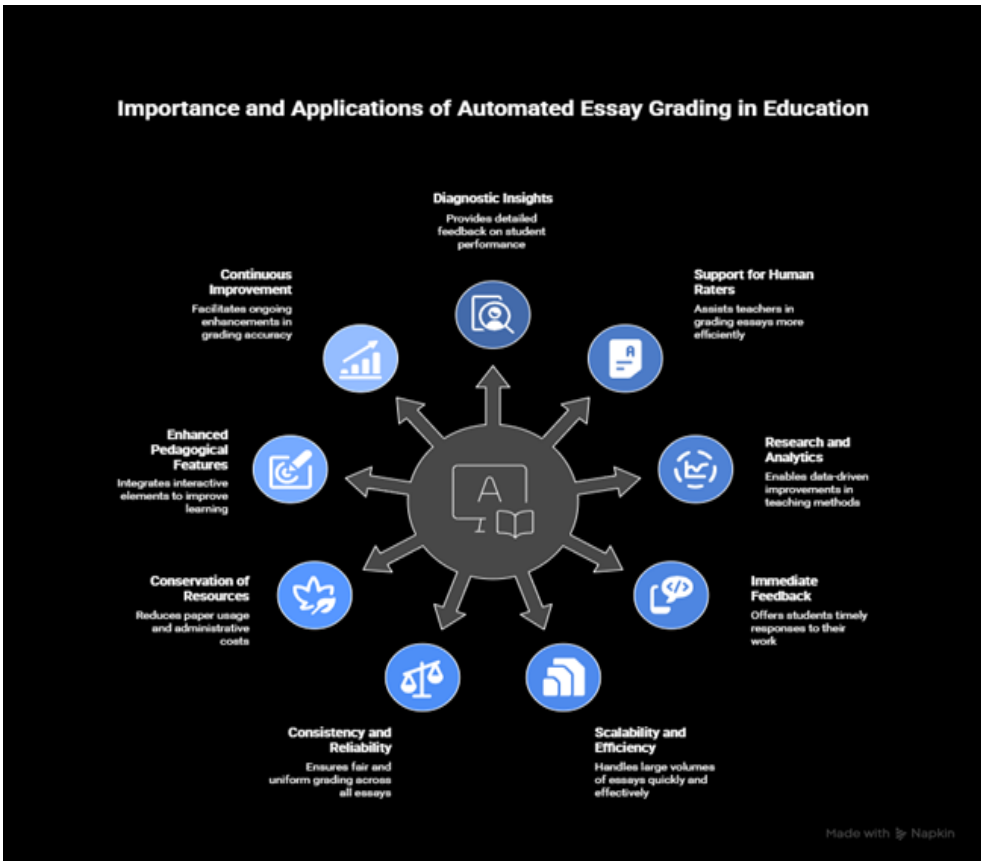


Figure 1. Importance and Applications of Automated Essay Grading in Education.

power, time, and paper-based materials. AEG systems can save a huge amount of human resources as well as prove to be an exceptional initiative towards sustainable lifestyle. Furthermore the development of different hybrid models and ensemble methods used in the field of AEG would also lead to advancement in educational research leading to development of novel techniques. AEG systems can be deployed in various datasets leading to improvement of their accuracy and reliability over time as they can adapt themselves according to the new dataset. This would lead to creation of more refined and sophisticated machine learning and deep learning models.

LITERATURE REVIEW

Historical Perspective

The timeline diagram of Automated Essay Grading depicts the sequential change of the algorithms along with the significant milestones (Figure 2). Early Automated Grading Systems (1960s-1980s): Ellis Page's pioneering work in the 1960s with the Project Essay Grade (PEG) system was the beginning of the timeline of automated essay grading which highlighted those 30 surface level features of essays-such as length, word frequency, punctuation patterns *etc*., could serve as proxies for underlying writing quality constructs. Correlations with human raters comparable to inter-rater reliability, establishing evidence of concept for computational essay assessment prove this work to be a significant milestone in this journey, (Ifenthaler D. *et al*., 2023)

Latent Semantic Analysis Era (1990s): The application of Latent Semantic Analysis (LSA), a technique for extracting semantic relationships from large text corpora was the next leap of this journey which overcame the major disadvantage of the earlier era reliance on only surface features. This Intelligent Essay Assessor (IEA) was developed by Thomas Landauer and his colleagues. This model could evaluate content relevance and conceptual coverage by comparing student essays to high-scoring exemplars or domain texts in semantic space (Shapiro, A. M., *et al*., 2000).

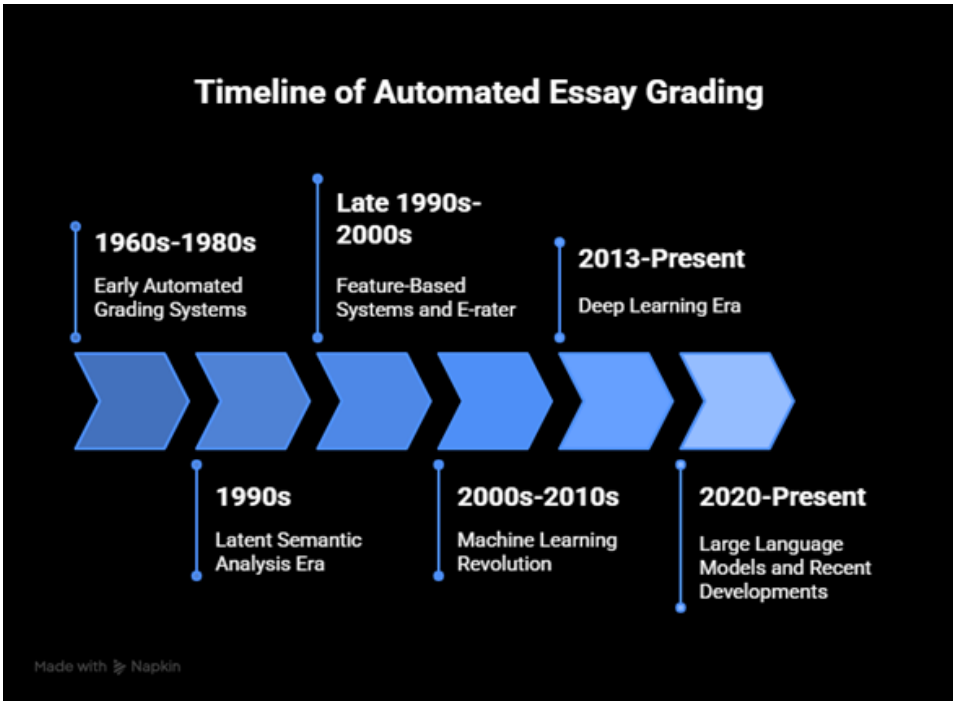


Figure 2. Timeline of Automated Essay Grading.

Feature-based Systems and E-rater (Late 1990s-2000s): In the late 1990s and 2000s the development of sophisticated feature-based systems especially E-rater (Electronic Essay Rater) by Educational Testing Service (ETS) proved to be a multimodal system which combined multiple feature categories. (Attali, Y., *et al.*, 2006) E-rater employed natural language processing techniques including part-of-speech tagging, syntactic parsing, and discourse analysis, combined with machine learning models to predict final scores. Many more notable systems like IntelliMetric, MY Access!and Criterion were also proposed during this era (Elliot S., *et al.*, 2003; Zwick R., *et al.*, 2012).

Machine Learning Revolution (2000s-2010s): Most of the machines learning algorithms were developed in the 2000s. Firstly Support Vector Machines (SVM) was deployed for classification and regression tasks in AEG due to their effectiveness of handling data with high-dimensional features. Random Forests also suggests promising results as they are particularly effective at handling the complex, non-linear relationships between features and scores. Feature extraction explosion of AEG also resulted during this era due to extensive research. Different features like lexical, syntactic, discourse, semantic and pragmatic were inspected and feature engineering combined with ensemble learning to achieve Quadratic Weighted Kappa (QWK) scores above 0.80 on several essay sets (Ramesh D., *et al.*, 2022 & Sharma S., *et al.*, 2020)

Deep Learning Era (2013-Present): The deep learning revolution transformed AEG emerging around 2013-2015. Word Embedding models like Word2Vec (Mikolov*et al.*, 2013) and GloVe (Pennington *et al.*, 2014) shows dense vector representations capturing semantic relationships, replacing sparse bag-of-words or one-hot encodings. Extensive studies in Recurrent Neural Networks such as LSTMs and GRUs exhibit the ability to model sequential dependencies in text, capturing long-range relationships between words and sentences. Convolutional Neural Networks on the other hand proved surprisingly effective for text classification tasks by learning local n-gram patterns. Hybrid CNN-LSTM models combined local pattern detection with sequential modelling to achieve higher accuracy. The attention mechanism (Chorowski *et al.*, 2015) enabled models to concentrate on relevant parts of input text, improving both performance and interpretability. ELMo (Peters *et al.*, 2018), ULMFiT (Howard & Ruder, 2018) and especially BERT (Liu*et al.*, 2019) demonstrated that pre-training on large general corpora followed by task-specific fine-tuning could achieve outstanding results. Transformer-based models BERT, RoBERTa, ALBERT and domain-specific variants consistently outperform previous approaches and prove that NLP models exhibits state of the art results.

Large Language Models and Recent Developments (2020-Present): Technological boom with Large Language Models (LLMs) such as GPT-3, GPT-4, ChatGPT, and open-source alternatives has opened new frontiers in this arena. In the era of rapidly changing assessment needs LLMs can evaluate essays with minimal or no task-specific training with the help of customised prompts according to the requirements. LLMs can also generate multi-dimensional detailed, personalised feedback and thus address one of the longstanding limitations of AEG systems (Atkinson *et al.*, & Liew *et al.*, 2024).

CONCLUSION

The journey from feature engineering to deep learning and LLMs of AEG has methodically unveiled the strengths and weaknesses of each model. For deployment and pedagogical purpose SHAP and rubric-level attributes provides interpretable scores in the scenarios where explainability matters. On the other hand attention and trait decomposition provide complementary, localised evidence for decisions. To create a rubric-decomposed, pedagogical grading system, combining trait-aware architectures with explainable attributions yields the most practically useful outputs where trait models shows interpretable sub-scores whereas SHAP or feature pools provide human-readable evidence for targeted feedback. For transfer and low-resource prompts, prompt-agnostic and trait-relation-aware models outperform prompt-specific systems because they learn generalised essay

representations instead of topic-bound features. For scalable deployment, pairwise comparative scoring and atomic decomposition produce more organised and calibrated outputs from LLMs and balancing compute and alignment considerations. Depending on the scenario when labelled data are very limited traditional feature-based and ensemble methods gives us competitive baselines and interpretability. On the other hand deeplearning models (BiLSTM/CNN/ transformer fine-tuning) improves representation learning and trait prediction when large amount of labelled data and linguistic features are available. Finally LLM-centric approaches mostly reduce labelling cost but require proper prompting, calibration, and often additional rubrics to match human standards.

REFERENCES

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports, 15*(1), 14551.

Attali, Y., & Burstein, J. (2006).Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *NIPS15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, 1, 577-585.

Elliot, S. (2003). Automated essay scoring: A cross-disciplinary perspective. *IntelliMetric: From here to validity*, p. 71-86.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1*. https://doi.org/10.18653/v1/P18-1031

Ifenthaler, D. (2023). Automated essay scoring systems. *Handbook of Open, Distance and Digital Education. Springer*, p. 1057-1071. https://doi.org/10.1007/978-981-19-2080-6_59

Liew, P. Y., & Tan, I. K. (2024).On automated essay grading using large language models. *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, p. 204-211.https://doi.org/10.1145/3709026.3709030

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). Roberta: A robustly optimized BERTpretraining approach. https://doi.org/10.48550/arXiv.1907.11692

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013).Distributed representations of words and phrases and their compositionality. *Computation and Language, 26*. https://doi.org/10.48550/arXiv.1310.4546

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* p. 1532-1543.https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation.*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, p. 1499-1509. https://doi.org/10.18653/v1/D18-1179

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Shapiro, A. M., & McNamara, D. S. (2000).The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research, 22*(1), 1-36. https://doi.org/10.2190/M811-G475-WKMX-X0JH

Sharma, S., & Goyal, A. (2020). Automated essay grading: An empirical analysis of ensemble learning techniques. *Computational Methods and Data Engineering: Proceedings of ICMDE. Springer, 2*, 343-362.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x